Zwischen Tradition und Moderne: Machine Learning in der Persönlichkeitsmessung

Ricarda Lübke

München 2020

Zwischen Tradition und Moderne: Machine Learning in der Persönlichkeitsmessung

Ricarda Lübke

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Fakultät für Psychologie und Pädagogik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Ricarda Lübke
geboren in Erding

München, den 07.10.2019

Ich habe immer geglaubt, und ich glaube immer noch, dass wir allem, ob es Glück oder Unglück bringt, dass wir allem was uns in den Weg kommt, insgesamt immer einen Sinn geben und es in etwas Wertvolles verwandeln können.

Hermann Hesse, Siddhartha

Erstgutachter: Prof. Dr. Markus Bühner

Zweitgutachter: Prof. Dr. Moritz Heene

Tag der mündlichen Prüfung: 05.03.2020

Vorwort

Diese Dissertation entstand im Rahmen meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Lehrstuhl für Psychologische Methodenlehre und Diagnostik der Ludwig-Maximilians-Universität München unter der Leitung von Prof. Dr. Markus Bühner.

An dieser Stelle möchte ich mich bei allen Personen, die auf unterschiedliche Weise zu dieser Dissertation beigetragen haben, herzlich bedanken. Besonderer Dank gilt:

- Markus Bühner für die Möglichkeit, mich immer weiterzuentwickeln und dafür, dass er mir immer den Rücken freigehalten hast.
- Moritz Heene für die vielen guten Gespräche bei einer Tasse Kaffee und die Bereitschaft, die Rolle als Zweitgutachter zu übernehmen.
- Sven Hilbert für die Bereitschaft, die Rolle als Drittprüfer zu übernehmen.
- Meinen Kollegen für das freundschaftliche und unterstützende Arbeitsumfeld sowie die vielen Diskussionen und Ideen, die wir ausgetauscht haben.
- David, Florian, Henning, Lena, Mirka, Philipp, Quay und Ramona fürs Korrekturlesen.
- Meinem Studenten Thomas für die fleißige Mitarbeit.
- Meiner Familie, Ingmar und allen anderen, die mir mich mit ihrem Rückhalt in der gesamten Zeit unterstützt haben.

In der vorliegenden Arbeit wird die männliche Form verwendet. Dies impliziert keine Geschlechtsdiskriminierung oder Verletzung des Gleichheitsgrundsatzes, sondern dient ausschließlich der sprachlichen Vereinfachung zur besseren Lesbarkeit.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CFA konfirmatorische Faktorenanalyse

EFA explorative Faktorenanalyse

LASSO Least absolute shrinkage and selection operator

ML Machine Learning

PCA Hauptkomponenten-Analyse (Principal Component Analysis)

RF Random Forest

SVM Support Vector Machine

Zusammenfassung

Machine Learning Algorithmen werden bereits in der sozialwissenschaftlichen Forschung als neue Methoden eingesetzt, um speziell entwickelte Fragestellungen zu beantworten. Darüber hinaus kann für solche neuen Methoden exploriert werden, inwiefern diese zur Beantwortung bestehender Fragestellung verwendet werden können. Hieran setzt die vorliegende Arbeit an, indem sie sich der Frage widmet, inwiefern Machine Learning in der psychologischen Fragebogentwicklung sinnvoll eingesetzt werden kann. Dazu wurden zwei Studien durchgeführt, welche an unterschiedlichen Aspekten des Fragebogenentwicklungsprozesses ansetzen. In beiden Studien wird der Fragebogenentwicklungsprozess im Rahmen der Persönlichkeitsmessung betrachtet.

In der ersten Studie wurde untersucht, inwiefern Machine Learning verwendet werden kann, um das Potenzial von Bildern als Stimulusmaterial für die Persönlichkeitsmessung abzuschätzen und erste Messmodelle abzuleiten. Hierzu wurden in Benchmark-Experimenten einerseits lineare (Multiple Regression, LASSO) und non-lineare (Support Vector Machine, Random Forest, Entscheidungsbaum) Modellierungen zur Vorhersage von Persönlichkeitseigenschaften verglichen. Andererseits wurde überprüft, inwiefern aus den Modellen mit der höchsten Vorhersagegüte Messmodelle für die jeweilige Persönlichkeitseigenschaft abgeleitet werden können.

Die zweite Studie befasst sich damit, inwiefern Machine Learning Algorithmen zusammen mit Panel Daten verwendet werden können, um Hinweise auf die Validität, insbesondere Inhaltsvalidität, von Fragebögen zu erhalten. Hierzu wurden ebenfalls Benchmark-Experimente durchgeführt, um die Vorhersagegüte linearer (LASSO) und non-linearer (Support Vector Machine, Random Forest) Modelle miteinander zu vergleichen. Darüber hinaus wurden die wichtigsten Prädiktoren aus den Modellen mit der höchsten Vorhersagegüte mit den Variablen mit den höchsten Korrelationen verglichen.

Die Ergebnisse der beiden Studien zeigen, dass durch die Verwendung von Machine Learning Algorithmen für den Fragebogenentwicklungsprozess relevante Informationen gewonnen werden können. Allerdings wird deutlich, dass sie als explorative Ansätze in jedem Fall von konfirmatorischen Analysen komplementiert werden

sollten. Darüber hinaus zeigen die Ergebnisse, dass die Zusammenhänge in den verwendeten Daten gut durch regularisierte lineare Modellierungen angenähert werden konnten.

Kapitel 1

Einleitung

Durch die rapide Entwicklung digitaler Systeme und die enorme Leistungssteigerung von Computern sind heutzutage viele statistische Methoden mit geringem Aufwand anwendbar, welche vor einigen Jahren noch Personen und Institutionen mit besonderen Rechenkapazitäten vorenthalten waren. So erleben Methoden, zu denen die Grundsätze bereits vor vielen Jahren entwickelt wurden, in den letzten Jahrzehnten einen Aufschwung und somit enorme Weiterentwicklung, da sie nun für viel mehr Forscher zugänglich sind. Beispiele für solche Methoden sind bayesianische Schätzmethoden und statistisches Lernen (auch Machine Learning, ML)[1]. Die Grundlagen bayesianischer Schätzmethoden wurden bereits 1763 veröffentlicht, wobei die Zahl der auf diesen theoretischen Annahmen basierten Publikationen erst seit den 1990er Jahren drastisch steigt (Corfield & Williamson, 2013). Statistisches Lernen wurde bereits in den 1960er Jahren theoretisch fundiert, aber findet erst seit Mitte der 1990er Jahre in praktischen Bereichen Anwendung (Vapnik, 1999).

ML beschreibt die Modellierung von Zusammenhängen zwischen Prädiktoren und Kriterien mit dem Ziel, auf Basis der Werte auf den Prädiktoren Vorhersagen für die Kriterien machen zu können, wenn dafür keine Daten vorliegen (James, Witten, Hastie & Tibshirani, 2013). Eine detaillierte Beschreibung ist Kapitel 1.3 zu entnehmen.

Während ML anfangs vor allem in anderen Bereichen angewendet wurde, ist es mittlerweile auch in der sozialwissenschaftlichen Forschung angekommen (Oswald & Putka, 2017; Yarkoni & Westfall, 2017) und stellt somit eine neue Methoden in diesem Forschungsfeld dar. Durch die Verfügbarkeit solcher neuer Methoden können einerseits neue Fragestellungen in einem Forschungsfeld entwickelt und beantwortet

[1] Im Folgenden werden die Begriffe statistisches Lernen und Machine Learning synonym verwendet.

werden. Andererseits kann exploriert werden, inwiefern diese Methoden verwendet werden können, um bereits vorhandene Fragestellungen zu beantworten.

Die Beantwortung neuer Fragestellungen mit Hilfe von ML kann in der Persönlichkeitspsychologie illustriert werden. Hier sind durch die Verfügbarkeit enormer Datenmengen aus digitalen Fußabdrücken von Personen, Fragestellungen dazu entstanden, wie aus diesen Fußabdrücken Persönlichkeitseigenschaften vorhergesagt werden können. Die Ergebnisse entsprechender Studien zeigen, dass aus verschiedenen digitalen Fußabdrücken wie Social Media Profilen und Aktivitäten (z. B. Golbeck, Robles & Turner, 2011; Kosinski, Stillwell & Graepel, 2013; Segalin et al., 2017a), aber auch digitaler Kommunikation (z. B. Boyd & Pennebaker, 2017; Peng, Liou, Chang & Lee, 2015; Quercia, Kosinski, Stillwell & Crowcroft, 2011) Persönlichkeitseigenschaften vorhergesagt werden können. Ohne den Einsatz von ML wären diese Fragestellungen nicht beantwortbar gewesen, da die Analyse der dahinter stehenden Datenmengen mit gängigen Methoden der psychologischen Forschung nur begrenzt möglich wäre.

Darüber hinaus gibt es erste Ansätze, um bereits vorhandene Fragestellungen, mit ML zu beantworten. Beispielsweise fordern Bleidorn, Hopwood und Wright (2017) dazu auf, ML auch in der Evaluation von Fragebögen einzusetzen. In einem weiteren Artikel beschreiben Bleidorn und Hopwood (2019), wie ML zur Weiterentwicklung von Persönlichkeitstheorie und -Messung verwendet werden kann, da in den entsprechenden Studien oftmals umfangreiche Datensätze verwendet werden, um Vorhersagen für Persönlichkeitseigenschaften zu machen. Somit birgt es die Möglichkeit, auch Zusammenhänge zwischen Variablen zu erkennen, die über die initiale theoretische Konzeption hinausgehen.

In der vorliegenden Arbeit wird auf Basis zweier Studien exploriert, inwiefern ML im Bereich der Fragebogenentwicklung eingesetzt werden kann. Die erste Studie befasst sich mit der Vorhersage von Persönlichkeitseigenschaften auf Basis von Bildbewertungen im Rahmen eines bildbasierten Fragebogens. Die Verwendung von Bildern anstelle textbasierter Items ist in der psychologischen Forschung bisher nur wenig verbreitet. Daher besteht die übergeordnete Fragestellung dieser Studie darin, zu explorieren, inwiefern ML dabei helfen kann, das Potenzial neuer, innovativer Verfahren zur Messung psychologischer Konstrukte abzuschätzen.

Die zweite Studie befasst sich mit der Vorhersage von Persönlichkeitseigenschaften auf Basis von Panel-Daten. Im Rahmen von Panel-Erhebungen werden viele Informationen über die Teilnehmer gesammelt. Die übergeordnete Fragestellung dieser Studie besteht darin, zu explorieren, inwiefern solche vorhandenen Datensätze

in Kombination mit ML zur Validierung von Fragebögen verwendet werden können.

Da sich beide Studien mit Persönlichkeitsmessung befassen, wird im Folgenden ein kurzer Überblick über die Persönlichkeitsforschung sowie die Fragebogenentwicklung gegeben. Anschließend werden die in der vorliegenden Arbeit verwendeten Methoden beschrieben. Darauf folgt die Darstellung der beiden empirischen Studien. Abschließend wird eine allgemeine Diskussion über die bearbeiteten Fragestellungen geführt.

1.1 Persönlichkeitseigenschaften

Bisherige Forschung zeigt, dass Persönlichkeitseigenschaften mit vielen verschiedenen Zielvariablen assoziiert sind (z. B. Ozer & Benet-Martínez, 2006), wie beispielsweise Lebenszufriedenheit (z. B. DeNeve & Cooper, 1998) und Arbeitserfolg (z. B. Barrick & Mount, 1991; Salgado, 1997). Daher ist zu erwarten, dass das Forschungsinteresse an Persönlichkeitseigenschaften auch zukünftig bestehen wird. Hierfür spricht auch, dass sich die Persönlichkeitspsychologie in vielen Anwendungsbereichen, wie beispielsweise der Adaption von Systemen an Anwender, zunehmender Beliebtheit erfreut (z. B. De Carolis & Mazzotta, 2011; Tkalcic, Kunaver, Kosir & Tasic, 2011). Da Persönlichkeitseigenschaften auch mit Ausbildungs- und Berufserfolg sowie Leistung zusammenhängen (z. B. Barrick & Mount, 1991; Judge, Higgins, Thoresen & Barrick, 1999; Tett, Jackson & Rothstein, 1991; Volodina, Nagy & Köller, 2015), werden sie auch in der Personalauswahl berücksichtigt (Morgeson et al., 2007).

Persönlichkeitseigenschaften werden allgemein als relativ stabil angesehen (z. B. Costa & McCrae, 1988; McCrae & Costa, 1982). Es wird davon ausgegangen, dass sie sich im Laufe des Lebens zwar verändern, allerdings mit zunehmendem Alter immer weniger Variabilität aufweisen (Roberts & DelVecchio, 2000). Die Rangreihe zwischen verschiedenen Personen bleibt allerdings weitgehend erhalten (Roberts & DelVecchio, 2000).

Big Five In den letzten Jahrzehnten hat sich maßgeblich das Konzept der Big Five in der psychologischen Forschung etabliert (McCrae & Costa, 2008). Dieses Konzept basiert auf einem lexikalischen Ansatz, bei dem davon ausgegangen wird, dass Eigenschaften, welche Persönlichkeit beschreiben, in der natürlichen Sprache durch Wörter repräsentiert sind (McCrae & John, 1992; McCrae & Costa, 1987). Obwohl das Konzept der Big Five Persönlichkeitseigenschaften im angloamerikani-

schen Raum entwickelt wurde, konnte das Vorhandensein dieses Konzepts auch in einer Vielzahl anderer Kulturen gezeigt werden (McCrae & Costa, 1997; Schmitt et al., 2007). Dabei ist darauf hinzuweisen, dass dies nicht bedeutet, dass die Entwicklung einer Klassifikation für Persönlichkeitseigenschaften basierend auf dem lexikalischen Ansatz in allen Kulturen das gleiche Ergebnis erzielt hätte (z. B. Cheung, van de Vijver & Leong, 2011; De Raad, Perugini, Hrebickova & Szarota, 1998).

Die fünf Persönlichkeitseigenschaften, die dem Big Five Konstrukt zugrunde liegen, sind Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit und Neurotizismus (McCrae & Costa, 1987). *Offenheit* beschreibt Personen, die eine lebendige Fantasie haben, Ästhetik schätzen, Gefühle intensiv erleben, experimentierfreudig und offen für Ideen sowie aufgeschlossen gegenüber Werte- und Normensystemen sind (Ostendorf & Angleitner, 2004). *Gewissenhaftigkeit* beschreibt Personen, die kompetent, ordnungsliebend, pflichtbewusst, selbstdiszipliniert und besonnen sind sowie ein hohes Leistungsstreben zeigen (Ostendorf & Angleitner, 2004). *Extraversion* beschreibt Personen, die herzlich, gesellig, durchsetzungsfähig und aktiv sind, Erlebnisse suchen sowie positive Emotionen wie Enthusiasmus erleben (Ostendorf & Angleitner, 2004). *Verträglichkeit* beschreibt Personen, die anderen vertrauen, unkompliziert, altruistisch, entgegenkommend, bescheiden und gutherzig sind (Ostendorf & Angleitner, 2004). *Neurotizismus* beschreibt Personen, die ängstlich, reizbar, depressiv, sozial befangen, impulsiv und verletzlich sind (Ostendorf & Angleitner, 2004).

1.1.1 Messung von Persönlichkeitseigenschaften

Zur Messung von Persönlichkeitseigenschaften können sowohl explizite als auch implizite Verfahren verwendet werden. Im Folgenden werden diese beiden Arten beschrieben.

Explizite Verfahren Die Messung von Persönlichkeitseigenschaften wird meist mit Fragebögen durchgeführt (Boyle, Matthews & Saklofske, 2008). Darin werden Personen klassischerweise Fragen zu typischem Verhalten gestellt, welche bezüglich des Zutreffens auf die eigene Person bewertet werden (Boyle et al., 2008; Cattell, 1946). Diese Art der Persönlichkeitserfassung hat sowohl Vor- als auch Nachteile (Boyle et al., 2008). Fragebögen stellen eine kostengünstige Art der Datenerhebung dar (Boyle et al., 2008) und dadurch, dass Fragebögen heutzutage oftmals online verbreitet und erhoben werden können, kann eine breite Masse erreicht werden (z. B. Krantz, Ballard & Scher, 1997; Miller et al., 2002; Riva, Teruzzi & Anolli, 2003). In Bezug auf die beiden Erhebungsmodi (offline und online) von Fragebögen konnte

gezeigt werden, dass die Ergebnisse von Befragungen keine systematischen Unterschiede zeigen (Krantz et al., 1997; Lonsdale, Hodge & Rose, 2006; Miller et al., 2002). Darüber hinaus konnte gezeigt werden, dass sich online und offline erhobene Stichproben bezüglich der psychometrischen Eigenschaften nicht signifikant voneinander unterscheiden (Lonsdale et al., 2006; Riva et al., 2003).

Beim Einsatz von Fragebögen ist allerdings zu beachten, dass das Ergebnis vor allem bei der Persönlichkeitsmessung relativ leicht verzerrt werden kann (z. B. Morgeson et al., 2007; Vecchione, Dentale, Alessandri & Barbaranelli, 2014). Einerseits spielen Verzerrungen durch soziale Erwünschtheit eine Rolle (Nederhof, 1985). Dabei stellen sich die Personen so dar, dass sie gesellschaftliche Normen erfüllen und somit als positiv wahrgenommen werden, unabhängig davon, wie sie wirklich sind (Nederhof, 1985). Andererseits kann es besonders in Auswahlsituationen passieren, dass Personen sich überlegen, welche Persönlichkeitseigenschaften für die angestrebte Stelle erwünscht sind und ihre Antworten entsprechend anpassen (z. B. Morgeson et al., 2007; Vecchione et al., 2014). Zudem kann kritisiert werden, dass klassische Persönlichkeitsfragebögen die Annahme zugrunde legen, dass das Selbst-Konzept von Personen bezüglich ihrer typischen Verhaltensweisen ihrem realen Verhalten entspricht (McCrae & Costa, 1982). Zusätzlich müssen sie in der Lage sein, über ihr Verhalten auf eine abstrakte Art und Weise zu reflektieren (Mackiewicz & Cieciuch, 2016).

Implizite Verfahren Neben Fragebögen zur expliziten Erfassung von Persönlichkeit, gibt es verschiedene Ansätze zur impliziten Erfassung (z. B. Egloff & Schmukle, 2002; Morgan & Murray, 1935; Rorschach, 1992). Darin schreiben Personen Geschichten zu Bildern, welche anschließend durch Experten evaluiert werden. Anhand der Experten-Evaluation können den Personen unterschiedliche Ausprägungen auf dem gemessenen Konstrukt zugeschrieben werden. Auch bei dieser Art der impliziten Erfassung können Verzerrungen im Antwortverhalten stattfinden (Vecchione et al., 2014).

1.2 Fragebögen in der psychologischen Forschung

Wie bereits in Kapitel 1.1.1 beschrieben, werden zur expliziten Messung von Persönlichkeitseigenschaften meist Fragebögen verwendet. Insgesamt stellen Fragebögen ein zentrales Instrument der psychologischen Forschung dar (Arnett, 2008). Im Folgenden soll ein kurzer Überblick zur Fragebogenentwicklung gegeben werden, da

in der vorliegenden Arbeit die Verwendung von ML in unterschiedlichen Bereichen der Fragebogenentwicklung exploriert wird. Diese Ausführung dient lediglich dem Verständnis und erhebt in keiner Weise den Anspruch, eine Anleitung zur Fragebogenentwicklung zu sein.

1.2.1 Fragebogenkonstruktion

Fragebögen werden in der psychologischen Forschung mit dem Ziel eingesetzt, Konstrukte valide zu messen (Clark & Watson, 1995).

Zu Beginn der Fragebogenentwicklung muss das zu untersuchende Konstrukt klar definiert werden (Clark & Watson, 1995). Hierbei ist es wichtig, eine theoretische Einbettung des Konstrukts in Bezug auf andere Konzepte zu erarbeiten (Clark & Watson, 1995). Anschließend werden Items formuliert, welche das definierte Ziel-Konstrukt vollumfänglich beschreiben, da im weiteren Verlauf der Fragebogenentwicklung lediglich nicht passende Items durch statistische Analysen identifiziert werden können, wohingegen Items, die zur Beschreibung des Konstrukts notwendig wären, nicht identifizierbar sind (Clark & Watson, 1995). Hierbei decken einzelne Fragen lediglich Teilaspekte des Konstrukts ab, sodass das gesamte Konstrukt durch eine Zusammensetzung verschiedener Items erfasst wird. Es sollte beachtet werden, dass die Teilaspekte des Konstrukts im finalen Fragebogen vollständig und ausgeglichen repräsentiert sind (Messick, 1995) bzw. die Repräsentation dem definierten Messmodell entspricht (Loevinger, 1957; Smith & McCarthy, 1995).

Die generierten Items sollten einerseits Standards der Itemformulierung entsprechen (Clark & Watson, 1995) und andererseits in der Lage sein, zwischen Personen mit unterschiedlichen Ausprägungen auf dem Merkmal zu unterscheiden (Borsboom, Mellenbergh & Van Heerden, 2004). Dabei sollte definiert sein, in welchen Bereichen ggf. eine besonders genaue Unterscheidung notwendig ist (Smith & McCarthy, 1995).

Nachdem mögliche Items generiert wurden und die Formulierungen mit qualitativen Methoden, wie beispielsweise kognitiven Interviews (Lenzner, Neuert & Otto, 2015), evaluiert und verbessert wurden, kann der Fragebogen einer quantitativen Evaluation unterzogen werden. Zu Beginn der quantitativen Evaluation sollte eine deskriptive Analyse der Itemantworten durchgeführt werden (Clark & Watson, 1995), um Unregelmäßigkeiten wie beispielsweise unbesetzte Antwortkategorien zu erkennen. Darüber hinaus werden unterschiedliche Gütekriterien für Fragebögen überprüft. Im Folgenden werden Objektivität, Reliabilität, Validität und Skalierung detaillierter beschrieben.

Objektivität Objektivität ist definiert als von der durchführenden Person unabhängige Durchführung, Auswertung und Interpretation (Eid & Schmidt, 2014).

Reliabilität Reliabilität ist als möglichst genaue, d.h. fehlerfreie, Messung des betrachteten Konstrukts definiert (Eid & Schmidt, 2014). Dabei wird zwischen der internen Konsistenz und Eindimensionalität unterschieden (Clark & Watson, 1995). Interne Konsistenz gibt an, wie hoch die Items untereinander korrelieren und steigt mit der Itemanzahl (Clark & Watson, 1995). Eindimensionalität hingegen gibt an, inwiefern die Items ein einzelnes Konstrukt messen, d. h. einen Faktor bilden (Clark & Watson, 1995).

Validität Validität ist so definiert, dass der Fragebogen ausschließlich das intendierte Konstrukt misst (Cattell, 1946) und eindeutig ist, was ein bestimmtes Messergebnis bedeutet (Messick, 1995). Hierbei wird zwischen verschiedenen Arten unterschieden (Messick, 1995). *Inhaltsvalidität* beschreibt, dass das Konstrukt in seiner gesamten Breite in dem Fragebogen repräsentiert ist und dieser somit alle inhaltlichen Aspekte gleichermaßen abbildet (Cronbach & Meehl, 1955; Smith & McCarthy, 1995). *Kriteriumsvalidität* beschreibt, dass der Zusammenhang zwischen dem gemessenen Konstrukt und einem manifesten Merkmal auch durch die Ergebnisse des Fragebogens reproduziert werden kann (Cronbach & Meehl, 1955). *Konstruktvalidität* beschreibt den Zusammenhang zwischen dem Fragebogen und anderen Konstrukten, welche indirekt z. B. durch Fragebögen gemessen wurden (Cronbach & Meehl, 1955). Hierbei wird zwischen diskriminanter und konvergenter Validität unterschieden. Diskriminante Validität bedeutet, dass das Ergebnis des Fragebogens keinen Zusammenhang zu einem Konstrukt aufweist, welches von dem gemessenen Konstrukt verschieden ist (Messick, 1995; Smith & McCarthy, 1995). Konvergente Validität hingegen bedeutet, dass das Ergebnis des Fragebogens mit dem Ergebnis einer anderen Methode zusammenhängt, die das gleiche Konstrukt messen soll (Messick, 1995). Konstruktvalidität setzt demnach voraus, dass eine theoretische Grundlage existiert, auf deren Basis Zusammenhänge mit anderen Konstrukten und Variablen abgeleitet werden können (Clark & Watson, 1995).

Skalierung Skalierung ist definiert als eine auf einem validen, psychometrischen Messmodell basierende Zuordnung von Antworten zu einem Messergebnis (Loevinger, 1957). Dieses Messmodell wird anhand der theoretischen Fundierung abgeleitet und seine Passung muss empirisch überprüft werden (Clark & Watson, 1995; Smith & McCarthy, 1995). Aus diesem Messmodell lässt sich zudem ableiten, wie die einzelnen Itemantworten verrechnet werden, um einen Gesamtwert für das jeweilige

Konstrukt zu erhalten (Smith & McCarthy, 1995). Um eine quantitative Evaluation eines Fragebogens zu ermöglichen, ist es notwendig, dass der neu entwickelte Fragebogen sowie Instrumente, die zur Validierung verwendet werden sollen, von einer großen und, bezüglich der Ausprägung auf dem betrachteten Konstrukt heterogenen, Stichprobe beantwortet wird (Clark & Watson, 1995). Die Heterogenität der Stichprobe ist insofern wichtig, als dass es lediglich möglich ist, Unterschiede zu messen, wenn solche auch in der Stichprobe vorliegen (Borsboom et al., 2004). Rouquette und Falissard (2011) zeigen in ihrer Simulationsstudie, dass die erforderliche Stichprobengröße zur Evaluation von Fragebögen abhängig von der Itemanzahl und der Anzahl der Dimensionen im Fragebogen ist und diese mit zunehmender Anzahl von Dimensionen und abnehmender Itemanzahl steigt.

Die Fragebogenentwicklung stellt einen iterativen Prozess dar, in welchem die Gütekriterien überprüft werden und anschließend auf Basis der Ergebnisse Veränderungen am Fragebogen vorgenommen werden (Clark & Watson, 1995; Smith & McCarthy, 1995). Bei der Anpassung des Fragebogens auf Basis der Evaluationsergebnisse können Konflikte zwischen verschiedenen Gütekriterien entstehen, sodass eine reflektierte Abwägung notwendig ist, bevor Items aus dem Pool entfernt werden (Clark & Watson, 1995). Ein häufig auftretender Konflikt zwischen den Gütekriterien kann zwischen Reliabilität und Validität bestehen. Dieser beschreibt, dass eine Selektion von Items basierend auf der Inter-Item-Korrelation dazu führen kann, dass die Validität verringert wird, da es hierdurch passieren kann, dass nur eine geringe Breite des Konstrukts erfasst wird (Clark & Watson, 1995; Cronbach & Meehl, 1955; Smith & McCarthy, 1995).

Für die resultierende neue Version des Fragebogens müssen anschließend wieder die Gütekriterien überprüft werden. Für jede Überprüfung wird dabei eine neue Stichprobe benötigt (Loevinger, 1957; Smith & McCarthy, 1995; Tuckey, 1950).

1.3 Computerintensive Modellierungsansätze

Aus der steigenden Zahl der Publikationen mit psychologischen bzw. sozialwissenschaftlichen Fragestellungen, in denen ML eingesetzt wird, kann gefolgert werden, dass die sozialwissenschaftliche Forschung in den letzten Jahren als Anwendungsfeld für ML entdeckt wurde (Oswald & Putka, 2017). Neben Artikeln, in denen die Anwendung solcher Methoden in den Verhaltenswissenschaften diskutiert wird (z. B. Bleidorn & Hopwood, 2019; Yarkoni & Westfall, 2017), erschienen auch Artikel, in denen ML Algorithmen angewandt wurden (z. B. Golbeck et al., 2011; Kosinski

et al., 2013).

Grundsätzlich wird im ML zwischen zwei grundlegenden Ansätzen unterschieden: überwachtem (engl. supervised Machine Learning) und unüberwachtem (engl. unsupervised Machine Learning) statistischen Lernen. Überwachtes Lernen zeichnet sich dadurch aus, dass die Variablen für die Modellierung in Prädiktoren und Zielvariable unterteilt werden und ein Modell erstellt wird, bei dem die Zielvariable durch die Prädiktoren vorhergesagt wird (z. B. Ghahramani & Jordan, 1993; Hastie, Tibshirani & Friedman, 2009; Jordan & Mitchell, 2015; Kotsiantis, 2007). Ziel einer solchen Modellierung ist es, ein Modell zu erstellen, welches für neue, unbekannte Beobachtungen, für die die jeweilige Ausprägung in der Zielvariable nicht vorliegt, auf Basis der vorhandenen Werte in den Prädiktoren eine möglichst genaue Vorhersage der Zielvariable vornimmt (z. B. Breiman, 2001b; Hastie et al., 2009).

Beim unüberwachtem Lernen hingegen ist keine Zielvariable vorhanden, sodass die eingesetzten Algorithmen auf Basis der Eigenschaften der Daten Ähnlichkeiten zwischen den Datenpunkten identifizieren (z. B. Hastie et al., 2009; Jordan & Mitchell, 2015; Kotsiantis, 2007).

Im Folgenden werden alle in der vorliegenden Arbeit verwendeten Algorithmen beschrieben. Dabei wird zunächst auf die verwendeten Methoden des überwachten Lernens und anschließend die Methoden des unüberwachten Lernens eingegangen.

1.3.1 Überwachtes Lernen

Den beiden empirischen Studien der vorliegenden Arbeit liegen Regressionsfragestellungen zugrunde. Dies bedeutet, dass die Zielvariable numerisch ist und als kontinuierlich angenommen wird. Andere Fragestellungsarten, wie Klassifikationen bei denen die Zielvariable kategorial ist, werden daher an dieser Stelle nicht weiter erläutert.

Im überwachten Lernen werden die Prädiktoren in einer n × p Matrix (X) dargestellt, wobei n die Anzahl der Personen und p die Anzahl der Prädiktoren darstellt (James et al., 2013). Ein einzelner Eintrag dieser Matrix wird mit x_{ij} notiert mit i $\in [1, n]$ und j $\in [1, p]$. Die beobachteten Werte auf der Zielvariablen werden in einem Vektor Y dargestellt, wobei einzelne Einträge als y_i notiert sind mit i $\in [1, n]$. Vorhergesagte Werte auf der Zielvariablen werden mit $\hat{y}_i$ notiert mit i $\in [1, n]$.

Kreuzvalidierung

Im Folgenden werden die Begriffe Trainings- und Testset verwendet. Als Trainingsset wird der Teil der Daten bezeichnet, anhand dessen die Modellparameter geschätzt werden und somit das Modell erstellt wird (Alexander, Tropsha & Winkler, 2015; James et al., 2013; Stone, 1974). Das Testset stellt einen davon unabhängigen Teil der Daten dar. Die Daten aus diesem Set werden in das zuvor erstellte Modell eingesetzt und somit ein vorhergesagter Wert bestimmt (Alexander et al., 2015; James et al., 2013; Stone, 1974). Anschließend wird die Vorhersagegüte des Modells ermittelt, indem der vorhergesagte Wert auf der Zielvariablen mit dem wahren Wert verglichen wird (Methoden zur Evaluation der Vorhersagegüte siehe Kapitel 1.3.1) (Alexander et al., 2015; James et al., 2013; Stone, 1974).

Bestimmt man die Vorhersagegüte im Trainingsset, d. h. In-Sample, wird diese meist überschätzt, da das Modell sich zu sehr an die Daten anpasst (Larson, 1931; Yin & Fan, 2001). Wird das Modell anschließend an vergleichbaren Daten überprüft, welche nicht zur Erstellung des Modells verwendet wurden (Out-Of-Sample), fällt die Vorhersagegüte schlechter aus als durch die In-sample Schätzung vermutet (Larson, 1931; Picard & Cook, 1984; Yin & Fan, 2001). Diese Differenz fällt meist größer aus, je mehr Variablen in dem Modell enthalten sind (Larson, 1931).

Da es Ziel der psychologischen Forschung sein sollte, robuste Zusammenhänge empirisch zu finden, was aufgrund der beschriebenen Probleme bei In-Sample Schätzungen nicht garantiert werden kann, fordert Haig (1996) zur Schätzung der Vorhersagegüte bei neuen Daten, Resampling-Methoden in der psychologischen Forschung einzusetzen. Besonders im Bereich des ML, wo oftmals große Zahlen an Prädiktoren einfließen, scheint dies besonders wichtig.

Im Bereich der Resampling-Methoden gibt es verschiedene Ansätze. Eine beliebte Methode ist die k-fache Kreuzvalidierung (Arlot & Celisse, 2010; Shmueli, 2010), welche im Folgenden detaillierter beschrieben wird.

Bei der k-fachen Kreuzvalidierung, wird der Datensatz zunächst zufällig in k ungefähr gleich große Teile aufgeteilt (Arlot & Celisse, 2010; James et al., 2013). Anschließend werden k-1 Teile als Trainingsdatensatz verwendet und das dabei entstandene Modell am übrigen Testdatensatz evaluiert (James et al., 2013; Picard & Cook, 1984). Der Vorgang von Training und Test wird k-mal wiederholt, sodass jeder der k Teile einmal als Testset fungiert und jeder Teil in k-1 Trainingssets verwendet wurde (Hastie et al., 2009). Hierbei ist darauf hinzuweisen, dass die Abschätzung der Vorhersagegüte nicht unbedingt mit einem größeren k steigt, da die Varianz größer wird, wenn $k \to n$, was besonders für Anwendungen suboptimal ist, in denen eine

Modell-Selektion stattfinden soll (Arlot & Celisse, 2010).

Um die Vorhersagegüte genauer zu bestimmen, werden für die Kreuzvalidierung oftmals mehrere Iterationen ($l > 1$) verwendet (Kim, 2009). In einer entsprechenden Simulationsstudie von Kim (2009) wurde gezeigt, dass sich die Ergebnisse von 10-fach wiederholter 10-facher Kreuzvalidierung bei Stichprobengrößen über 140 Personen als stabil erwiesen hat.

Viele Modelle, die im ML angewendet werden, verfügen über Hyperparameter (Probst, Bischl & Boulesteix, 2018). Dies sind Parameter, die nicht wie die anderen Modellparameter direkt durch den Schätzalgorithmus anhand der Daten geschätzt werden können, im Einzelfall jedoch sinnvoll gewählt werden müssen, damit das Modell eine optimale Vorhersage macht (Probst et al., 2018). Um diese Wahl zu treffen, wird daher oftmals zusätzlich zu der oben beschriebenen äußeren Schleife der Kreuzvalidierung noch eine innere Schleife hinzugefügt. Dies bedeutet, dass in jedem Schritt der äußeren Kreuzvalidierung das Trainingsset erneut in m Teile aufgeteilt wird, um anhand verschiedener Kombinationen der Hyperparameter diejenigen zu bestimmen, die eine optimale Vorhersage ermöglichen. Die so bestimmten Parameter werden anschließend in der jeweiligen äußeren Schleife bei der Schätzung der Vorhersagegüte verwendet.

Evaluation der Vorhersagegüte

Bei Methoden des überwachten Lernens kann die Vorhersagegüte bestimmt werden, indem die durch das Modell vorhergesagten Werte auf der Zielvariablen mit den tatsächlich beobachteten Werten auf der Zielvariablen verglichen werden. Für diesen Vergleich können unterschiedliche Herangehensweisen und Maße verwendet werden. Im Folgenden werden die in der vorliegenden Arbeit verwendeten Maße vorgestellt.

Es sei darauf hingewiesen, dass durch die Evaluation der Vorhersagegüte verschiedener Modelle nicht ermittelt werden kann, welche Art von Modell wirklich hinter den Daten steht (James et al., 2013). Durch die Evaluation kann lediglich überprüft werden, wie gut ein bestimmtes Modell die Zusammenhänge annähern kann (James et al., 2013).

Mean Squared Error (MSE) Im Bereich der Regressionsfragestellungen ist das meistgenutzte Maß der Mean Squared Error (MSE), welches sich wie in Formel 1.1 berechnet (James et al., 2013).

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1.1}$$

Hierbei wird für jede Beobachtung die quadrierte Differenz zwischen dem beobachteten Wert und dem durch das entsprechende Modell vorhergesagte Wert berechnet (RSS, siehe Formel 1.4), über alle Beobachtungen summiert und durch die Anzahl der Beobachtungen dividiert (James et al., 2013). Der Wertebereich des MSE liegt zwischen 0 und ∞, wobei ein Wert von 0 eine genaue Übereinstimmung zwischen vorhergesagten und beobachteten Werten bedeutet (James et al., 2013). Aufgrund der geringen Komplexität ist der MSE ist ein einfach verständliches und gut interpretierbares Maß, da es das Verhältnis aus durch das Modell erklärbarer Varianz und Rauschen darstellt (Wang & Bovik, 2009).

R_C^2 Bei Betrachtung des MSE können lediglich relative Aussagen über die Vorhersagegüte eines bestimmten Modells im Vergleich zu anderen Modellen gemacht werden. Ein Maß, welches eine absolute Einschätzung der Vorhersagegüte ermöglicht, ist das R_C^2. In der psychologischen Forschung ist R^2 aus linearen Regressionen bekannt und beschreibt die aufgeklärte Varianz (Ozer, 1985). Diese berechnet sich wie in Formel 1.2 dargestellt (James et al., 2013).

$$R_C^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.2}$$

wobei $\bar{y}$ der Mittelwert der Stichprobe (Alexander et al., 2015).

Somit stellt R^2 den Quotienten aus der Summe der quadrierten Fehler des Modells (RSS, siehe Formel 1.4) und der Summe des quadrierten Fehlers eines Modells ohne Prädiktorvariable (Null-Modell) dar (Alexander et al., 2015; Dougherty, Kim & Chen, 2000; Edwards, Muller, Wolfinger, Qaqish & Schabenberger, 2008; Magee, 1990). Der Wertebereich liegt demnach zwischen 0 und 1, wobei 1 für eine perfekte Beschreibung der Daten durch das Modell steht (Alexander et al., 2015; Kvalseth, 1985). Ein Wert von 0 hingegen liegt vor, wenn $\forall i\ \hat{y}_i = \bar{y}$, was bedeutet, dass der Fehler des Modells ebenso hoch ist wie der Fehler in einem reinen Intercept-Modell, in dem keine Prädiktoren enthalten sind, bzw. alle beta-Koeffizienten gleich 0 sind (Alexander et al., 2015).

Im Bereich des ML mit Resampling (vgl. Kapitel 1.3.1) hingegen weichen Wertebereich und Interpretation von dem zuvor beschriebenen Konzept für eine Stichprobe ab. R_C^2 stellt hier ein normiertes Maß der Vorhersagegüte dar, welches einen Wertebereich zwischen $-\infty$ und 1 hat. Ein Wert von 1 bedeutet eine perfekte Vorhersage

der gemessenen Zielvariable durch das Modell, während ein Wert ≤ 0 bedeutet, dass die Zielvariable nicht besser vorhergesagt werden kann als wenn das Modell den Mittelwert der Zielvariablen im Testset vorhersagt (Pargent & Gönna, 2018). Der negative Wertebereich kommt dadurch zustande, dass die quadrierte Summe der Fehler des im Trainingsset erstellten Modells um ein Vielfaches kleiner sein kann als die quadrierte Summe der Fehler in einem Modell ohne Prädiktoren im Testset (Alexander et al., 2015). Somit beschreibt das im Trainingsset erstellte Modell die Daten im Testset überhaupt nicht (Alexander et al., 2015).

Da das R^2_C durch die Normierung des Vorhersagefehlers einfacher zu interpretieren ist und R^2 ein in der Psychologie gängiges Maß darstellt, wurden alle Modelle in der vorliegenden Arbeit auf Basis dieses Maßes optimiert und abschließend miteinander verglichen.

Spearman-Rho Spearman-Rho stellt eine Rangkorrelation dar. Diese basiert auf der Annahme, dass nicht der absolute Wert, sondern nur die Rangreihe der Beobachtungen, von Interesse ist (Spearman, 1904). Dieses Maß hat außerdem den Vorteil, dass Ausreißer einen geringeren Einfluss haben, weshalb es als robustes Maß gilt (Rosset, Perlich & Zadrozny, 2007; Spearman, 1904). Darüber hinaus können durch Rangkorrelationen Variablen mit unterschiedlichen Verteilungen miteinander verglichen werden (Spearman, 1904).

Um Spearman-Rho zu berechnen, wird nach Transformation der betrachteten Variablen in Ränge, eine Pearson-Korrelation berechnet (Spearman, 1904). Daher liegt der Wertebereich des Spearman-Rho zwischen -1 und 1, wobei Werte von 1 einen perfekten positiven Zusammenhang, Werte von -1 einen perfekten negativen Zusammenhang und Werte von 0 keinen Zusammenhang beschreiben (Rosset et al., 2007).

Im Rahmen von Regressionsfragestellungen scheint es oftmals vor allem relevant, Personen mit einer hohen Ausprägung auf der Zielvariablen von denen mit einer niedrigen Ausprägung zu trennen (Rosset et al., 2007). Daher wird dieses Maß oft im ML verwendet und evaluiert den Zusammenhang zwischen den in Ränge transformierten, durch das Modell vorhergesagten Werten und den beobachteten Werten.

Rechenzeit Da Algorithmen neben den genannten Maße der Vorhersagegüte auch an ihrer Effizienz evaluiert werden (Ottmann & Widmayer, 2017), wurde in der vorliegenden Arbeit bei allen Algorithmen auch die Rechenzeit dokumentiert. Allerdings

wird diese nur rein deskriptiv aufgeführt und nicht weiter analysiert.

Zielkonflikt zwischen Verzerrung und Varianz (Bias-Variance Trade-Off)

Beim Vergleich verschieden komplexer Modelle ist darauf hinzuweisen, dass man sich in einem Zielkonflikt zwischen Verzerrung (Bias) und Varianz bewegt, da diese sich gegenläufig beeinflussen. Oftmals können komplexe Modelle die Daten im Trainingsset mit geringer Verzerrung beschreiben und führen dadurch im Trainingsset zu einem geringeren Vorhersagefehler (Hastie et al., 2009). Dieser Zusammenhang ist allerdings nicht auf die Daten im Testset übertragbar. Durch komplexe Modellierung kann es vorkommen, dass das Modell zu sehr an die Daten im Trainingsset angepasst ist, was zu einem hohen Vorhersagefehler im Testset führt (Hastie et al., 2009). Dies liegt an der hohen Varianz der Vorhersagen. Damit ist gemeint, dass sich Vorhersagen für neue Daten bei Verwendung unterschiedlicher Trainingssets stark voneinander unterschieden (Hastie et al., 2009). Durch sehr einfache Modellierung kann es hingegen passieren, dass das Modell die Daten im Trainingsset nicht gut genug beschreibt (Hastie et al., 2009). Auch dies führt im Testset zu hohen Vorhersagefehlern und das obwohl die Varianz klein ist, d.h. die Vorhersagen für neue Daten sehr robust gegenüber der Wahl des verwendeten Traingssets sind (Hastie et al., 2009). Aufgrund dieses Zielkonflikts wird versucht, Modelle so zu optimieren, dass sie einen möglichst geringen Vorhersagefehler haben und somit ein optimales Verhältnis zwischen Verzerrung und Varianz gefunden wird (Hastie et al., 2009).

Modellierung linearer Zusammenhänge

Lineare Modelle gibt es schon sehr lange, da sie vor allem bei einer geringen Anzahl von Prädiktoren ohne großen Rechenaufwand geschätzt werden können. Ein großer Vorteil linearer Modelle ist, dass die Ergebnisse gut interpretierbar sind (James et al., 2013). Wie der Name schon sagt, wird bei diesen Modellen der Zusammenhang zwischen den Prädiktoren und der Zielvariable linear modelliert. Hierbei wird angenommen, dass der zugrundeliegende Zusammenhang entweder linear ist oder durch ein lineares Modell approximiert werden kann (Hastie et al., 2009). Diese Annahme steht hinter einer Vielzahl von Analysen in der psychologischen Forschung.

Im Folgenden werden zwei lineare Modellierungsansätze vorgestellt: multiple lineare und regularisierte Regressionsmodelle.

Multiples Lineares Modell Sind in einem Modell mehrere Prädiktoren enthalten, so werden multiple lineare Regressionsmodelle verwendet. Diese können als eine

der Standardmethoden in der Psychologie gesehen werden, da sie sehr häufig eingesetzt werden (Nathans, Oswald & Nimon, 2012; Yin & Fan, 2001). Die generelle Formel für multiple lineare Modelle ist Formel 1.3 zu entnehmen (Kvalseth, 1985).

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon \tag{1.3}$$

wobei β_0 den Intercept (Konstante), X_j den j-ten Prädiktor, β_j den Zusammenhang zwischen dem j-ten Prädiktor und dem Kriterium (Konstante) und ϵ den Messfehler repräsentiert.

Im Rahmen des ML werden zur Vorhersage einer Zielvariablen im Trainingsset die Regressionsgewichte β so bestimmt, dass die Summe der kleinsten Quadrate (Residual Sum of Squares, RSS) minimiert wird (siehe Formel 1.4) (Hastie et al., 2009). Die hierdurch enthaltene Formel wird anschließend im Testset verwendet, sodass für jede Person im Testset die Beobachtungen auf den einzelnen Prädiktoren in die Formel eingesetzt werden und so ein vorhergesagter Wert auf der Zielvariablen ermittelt wird (James et al., 2013). Zur Bestimmung der Vorhersagegüte wird anschließend die Differenz zwischen dem vorhergesagten Wert auf der Zielvariablen und dem beobachteten Wert auf der Zielvariablen bestimmt (James et al., 2013) (verschiedene Maße der Vorhersagegüte sind in Kapitel 1.3.1 beschrieben).

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \sum_{j=1}^{N} x_{ij} \hat{\beta}_j)^2 \tag{1.4}$$

wobei $\hat{\beta}_0$, $\hat{\beta}_j$ die Schätzwerte der Regressionskoeffizienten sind, bei denen RSS so klein wie möglich ist (James et al., 2013).

LASSO (Least absolute shrinkage and selection operator) Bei einer großen Zahl an Prädiktoren ist die Interpretierbarkeit eines multiplen linearen Modells, welches alle Prädiktoren enthält, limitiert. Zudem können die Vorhersagen durch irrelevante Prädiktoren verschlechtert werden, da sich das Modell im Trainingsset zu sehr an die Daten anpasst (Hastie et al., 2009). Daher wurden Methoden zur Regularisierung linearer Modelle entwickelt. Eine solche Methode stellt das LASSO (Tibshirani, 1996) dar. Dieses führt gleichzeitig zur Modellierung des Zusammenhangs mit der Zielvariablen eine Reduktion der Anzahl der Prädiktoren durch (James et al., 2013). Zur Modellierung des Zusammenhangs wird auch beim LASSO eine lineare Regression (Formel 1.3) verwendet. Bei der Schätzung der Regressionsgewichte wird allerdings zusätzlich zu der Summe der kleinsten Quadrate ein Strafterm in Form einer ℓ_1 Norm eingeführt, welcher dazu führt, dass eher kleine Koeffizienten mit 0 geschätzt werden (Hastie et al., 2009). Formal wird daher der folgende Ausdruck

minimiert:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.5}$$

wobei λ als Strafterm dient und β_j das Regressionsgewicht für Prädiktor j ist (James et al., 2013).

Das λ stellt hier einen Hyperparameter dar, dessen optimaler Wert durch innere Kreuzvalidierung bestimmt wird.

Modellierung nicht-linearer Zusammenhänge

Zusammenhänge zwischen Prädiktoren und Zielvariablen können auch nicht-linear sein. Im Folgenden werden die in der vorliegenden Arbeit verwendeten Modellierungen (Support Vector Machine, Entscheidungsbaum und Random Forest) beschrieben.

Support Vector Machine (SVM) In Regressionsfragestellungen kann die SVM als nicht lineare-Regularisierung eines linearen Regressionsmodells beschrieben werden (Smola & Schölkopf, 2004). Diese legt um die Regressionsgerade herum einen Toleranz-Bereich (Drucker, Burges, Kaufmann, Smola & Vapnik, 1996; Smola & Schölkopf, 2004). Liegen Werte in diesem Bereich, so gehen sie nicht als Fehler ins Modell ein (Drucker et al., 1996; Smola & Schölkopf, 2004). Die Abweichung von der Regressionsgeraden der Werte, die außerhalb dieses Bereichs liegen, geht wie gewohnt in die Verlustfunktion zur Vorhersagegüte ein, wobei der Abstand zwischen Punkt und Toleranz-Bereich relevant ist (Drucker et al., 1996; Smola & Schölkopf, 2004). Um den Vorhersagefehler zu minimieren, wird im Rahmen innerer Kreuzvalidierung ein optimaler Toleranz-Bereich bestimmt (Smola & Schölkopf, 2004; Steinwart & Thomann, 2017). Durch die Einführung des Toleranzbereichs werden die beobachteten Werte in veränderten Vektoren abgebildet (Smola & Schölkopf, 2004). Anschließend folgt die Einführung eines Kernels, welcher die Abbildung der Datenpunkte in einen hoch-dimensionalen Raum vornimmt (Karatzoglou, Smola & Hornik, 2004), sodass die SVM die Daten nicht-linear modelliert (Smola & Schölkopf, 2004). Der Kernel ist eine Funktion, welche die Ähnlichkeit zweier Beobachtungen durch gewichtete Skalar-Produkte beschreibt (James et al., 2013; Smola & Schölkopf, 2004). Aus diesen Skalar-Produkten wird schließlich die Zielvariable vorhergesagt (Smola & Schölkopf, 2004). Da dies zu vielen verschiedenen Funktionen führen kann, wird abschließend die Funktion gewählt, welche die geringste Steigung aufweist (Smola & Schölkopf, 2004).

Formal ergibt sich die in der SVM verwendete Funktion zur Vorhersage neuer

Daten durch:

$$f(D, \lambda, \gamma) = arg \min_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + \frac{1}{n} \sum_{i=1}^{n} L_\omega(y_i, f(x_i)) \tag{1.6}$$

mit $\lambda > 0$ als Regularisierungs-Parameter, H_γ einem reproduzierbaren Kernel im Hilbertraum mit Kernel k_λ und dem Kernel-Parameter $\lambda > 0$, D einem Datensatz mit Prädiktor und Zielvariablen und L_ω einer Verlustfunktion mit Gewichtungs-Parameter $\omega > 0$ (Smola & Schölkopf, 2004).

Entscheidungsbaum (Regressionen) Ein Entscheidungsbaum ist eine Methode, bei der der Datensatz in verschiedene Teile aufgesplittet wird. Der Baum beginnt an der Wurzel und wächst anschließend nach unten, wobei an jeder Stelle, an welcher der Baum gesplittet wird, ein Knoten entsteht.

Im Folgenden wird aufgrund der Fragestellungen in der vorliegenden Arbeit ein Regressions-Baum beschrieben.

Um einen Regressions-Baum zu erstellen wird wie folgt vorgegangen (James et al., 2013; Therneau & Atkinson, 2018b):

1. Teile den Datensatz in zwei Teile und erstelle somit einen linken und rechten Teilbaum. Als Kriterium zum Teilen überprüfe, anhand welcher Variable der Datensatz am besten aufgeteilt werden kann, d.h. minimiere den Term in Formel 1.7 (Therneau & Atkinson, 2018b).

$$RSS_{Baum} - (RSS_{LinkerBaum} + RSS_{RechterBaum}) \tag{1.7}$$

   wobei für jeden (Teil-)Baum gilt $RSS_{Baum} = \sum_{j=1}^{N} (y_j - \bar{y})^2$

2. Wiederhole das Vorgehen aus 1. für jeden entstandenen Teilbaum so lange, bis ein Stopp-Kriterium erreicht wurde (Therneau & Atkinson, 2018b)[2].

3. Schneide den entstandenen Baum mit Hilfe des Komplexitätsparameters cp wieder zurück. Dabei liegt dieser Parameter zwischen 0 und ∞ und gibt an, wie sehr sich die Vorhersagegüte des Modells verbessern muss, damit das Hinzufügen einer weiteren Variable in einem Split umgesetzt wird (Therneau & Atkinson, 2018b). Bei einem Wert von $cp = 1$ werden die Daten nicht gesplittet (Therneau & Atkinson, 2018b).

[2] In der vorliegenden Arbeit wurden die folgenden Stopp-Kriterien angewendet: minbucket (Minimale Anzahl an Beobachtungen in einem Blatt), minsplit (Minimale Anzahl an Beobachtungen in einem Knoten, welcher weiter gesplittet werden darf), maxdepth (maximale Tiefe des Baums).

Ist der Entscheidungsbaum fertig aufgestellt, wird für alle Beobachtungen, die dem gleichen Blatt zugeordnet werden können, der Mittelwert aller Daten in diesem Blatt auf der Zielvariablen vorhergesagt (James et al., 2013).

Random Forest (RF) Ein RF ist eine Methode, welche auf Entscheidungsbäumen aufbaut. Dieser Algorithmus unterscheidet sich allerdings in einigen wesentlichen Punkten von einfachen Entscheidungsbäumen:

1. **Anzahl der Bäume:** Während bei einfachen Entscheidungsbäumen lediglich ein Baum aufgestellt und interpretiert wird, werden in einem RF mehrere Bäume anhand von Bootstrap-Stichproben des Trainingsdatensatzes geschätzt. Zur Berechnung der RF Vorhersage für eine neue Beobachtung werden die Vorhersagen aller Bäume gemittelt.
2. **Verwendete Prädiktoren:** Während bei einfachen Entscheidungsbäumen an jedem Knoten alle Prädiktoren für einen Split zur Verfügung stehen, steht bei einem RF lediglich eine Zufallsauswahl der Prädiktoren an jedem Knoten zur Verfügung. Die Vektoren mit den ausgewählten Prädiktoren an jedem Knoten sind voneinander unabhängig (Breiman, 2001a) und die Anzahl der zufällig ausgewählten Prädiktoren wird im R-Paket ranger durch die Variable mtry festgelegt (Wright & Ziegler, 2017). Durch die zufällige Entfernung der dominanten Variablen wird die Korrelation zwischen den verschiedenen Entscheidungsbäumen reduziert, was zu einer effektiveren Aggregation der Vorhersagen und damit in der Regel zu einer Verbesserung der Vorhersagegüte des resultierenden RF führt.

Ebenso wie im einfachen Entscheidungsbaum kann im RF festgelegt werden, wie viele Beobachtungen in einem Knoten mindestens vorliegen müssen, damit potenziell ein weiterer Split durchgeführt wird (Wright & Ziegler, 2017).

Interpretierbarkeit

Die oben beschriebenen Modelle unterscheiden sich in ihrer Interpretierbarkeit. Lineare Modelle wie multiple lineare Regressionen und LASSO, weisen eine hohe Interpretierbarkeit auf, da sie auf einfachen statistischen Methoden beruhen, bei denen die Regressionsgewichte einen Zusammenhang zwischen Prädiktoren und Kriterium angeben, der dem Muster „je ..., desto ...“ entspricht. Aber auch Entscheidungsbäume und hier vor allem Bäume, die lediglich eine geringe Tiefe aufweisen, gelten als einfach interpretierbar, da das erstellte Modell durch einen Menschen leicht ver-

standen werden kann und dies sogar eine intuitive Darstellung von Interaktionen beinhaltet.

Schwer interpretierbare Modelle sind hingegen SVM und RF. Um eine Interpretierbarkeit dieser Modelle zu ermöglichen, können nach dem Anpassen des Modells zusätzliche Methoden angewandt werden, die die Interpretierbarkeit erhöhen (Molnar, 2019). Im Folgenden werden die Methoden vorgestellt, die in der vorliegenden Arbeit verwendet wurden.

Permutation Feature Importance Die Permutation Feature Importance gibt an, wie sehr sich die Vorhersagegüte verschlechtert, wenn die Werte des betrachteten Prädiktors so verändert werden, dass sie nicht mehr den ursprünglichen Zusammenhang zur Zielvariable aufweisen (Fisher, Rudin & Dominici, 2018; Molnar, 2019). Um dies zu berechnen, werden die Werte eines Prädiktors über alle Beobachtungen hinweg permutiert und die dadurch entstehende Differenz in der Vorhersagegüte zwischen den ursprünglichen und den permutierten Daten berechnet (Molnar, 2019).

Global Surrogate Model Durch ein Global Surrogate Model wird versucht, mit Hilfe eines einfach zu interpretierenden Modells eine Annäherung an den durch den ausgewählten ML Algorithmus modellierten Zusammenhang zu finden (Molnar, 2019). Als Datenbasis zur Erstellung des Global Surrogate Modells dienen die Prädiktoren und die durch den gewählten ML Algorithmus vorhergesagten Werte. Um den Zusammenhang zwischen diesen Variablen zu modellieren werden gut interpretierbare Modelle wie beispielsweise ein lineares Modell mit wenigen Prädiktoren oder ein Entscheidungsbaum verwendet (Molnar, 2019).

Hierdurch können einerseits wichtige Prädiktoren identifiziert werden. Andererseits kann durch Betrachtung der Varianzaufklärung des Global Surrogate Modells angegeben werden, wie gut ein solches Modell die Ergebnisse des komplexeren Algorithmus approximiert (Molnar, 2019).

1.3.2 Unüberwachtes Lernen

Wie oben beschrieben (siehe Kapitel 1.3.1), werden in der vorliegenden Arbeit Regressions-Fragestellungen beantwortet. Da der Datensatz in Studie II jedoch viele Prädiktoren enthält, wurde zur Variablenselektion zunächst eine Dimensionsreduktion durchgeführt.

Die dazu verwendeten Methoden des unüberwachten Lernens werden im Folgenden beschrieben.

Hauptkomponenten-Analyse (PCA)

Ziel einer PCA ist es, relevante Informationen aus einem Datensatz zu extrahieren (Abdi & Williams, 2010; Jolliffe, 2002). Dafür werden korrelierte Variablen durch Singulärwertzerlegung in Dimensionen zusammengefasst, welche so festgelegt werden, dass die endgültigen Dimensionen orthogonal zueinander sind (Abdi & Williams, 2010; James et al., 2013). Die Anzahl der Dimensionen ist vor Durchführung der PCA festzulegen, wobei dafür unterschiedliche Methoden verwendet werden können (Abdi & Williams, 2010; Jolliffe, 2002).

Nach Abschluss der PCA enthält die Datenmatrix für jede Beobachtung und jede Dimension einen Wert, welcher als Faktorwert angesehen werden kann (Abdi & Williams, 2010). Der Eigenwert einer Dimension ist durch die Summe der quadrierten Faktorladungen auf dieser Dimension definiert (Abdi & Williams, 2010). Die Wichtigkeit einer Beobachtung für eine Dimension kann daher durch das Verhältnis aus dem quadrierten Fakorwert auf dieser Dimension und dem Eigenwert der Dimension berechnet werden (Abdi & Williams, 2010).

Clustering von Variablen

Clustering ist eine Methode zur Einteilung von Daten in Gruppen (Chavent, Kuentz-Simonet, Liquet & Saracco, 2012; Hastie et al., 2009; Vigneau & Qannari, 2003). Ziel ist es, Gruppen zu identifizieren, bei denen die Datenpunkte innerhalb einer Gruppe eine möglichst hohe Homogenität aufweisen, während eine maximale Unterscheidung zu anderen Gruppen angestrebt wird (Chavent et al., 2012; Hastie et al., 2009). In der vorliegenden Arbeit werden Variablen geclustert, sodass aus einer hohen Ähnlichkeit innerhalb einer Gruppe geschlussfolgert werden kann, dass alle darin enthaltenen Variablen eine ähnliche Information messen (Chavent et al., 2012). Hierbei ist hervorzuheben, dass mit Clustering keine Beschreibung der einzelnen Cluster einhergeht, sondern lediglich eine datengetriebene Einteilung in Gruppen realisiert wird (Jain, 2008).

Da die Ähnlichkeit zwischen den Gruppen im Clustering zentral ist, muss dem Algorithmus eine Funktion zur Beschreibung dieser Ähnlichkeit übergeben werden. Je nach verwendetem Algorithmus werden dafür verschiedene Maße verwendet (Jain, Murty & Flynn, 1999). In der vorliegenden Arbeit wurde der K-Means Algorithmus verwendet, welcher im Folgenden beschrieben wird.

K-Means Clustering K-Means Clustering gehört zu den Partitionierungs-Algorithmen, bei denen die Ähnlichkeit innerhalb der Cluster anhand der Methode der kleinsten Quadrate (siehe Formel 1.4) bestimmt wird (Jain et al., 1999). Da die

Anzahl der Cluster vor Verwendung des Algorithmus definiert werden muss, ist es eine top-down Clustering-Methode (Hartigan & Wong, 1979; Hastie et al., 2009).

Im Folgenden wird das Vorgehen des K-Means Algorithmus beschrieben (Jain et al., 1999):

1. Definiere ein Konvergenzkriterium.
2. Für die zuvor festgelegte Zahl k werden zufällig Punkte im p-dimensionalen Raum festgelegt, welche jeweils den Mittelpunkt eines Clusters repräsentieren.
3. Für jeden Datenpunkt wird die Distanz zu den zuvor festgelegten Mittelpunkten bestimmt und anschließend wird jeder Datenpunkt dem Mittelpunkt eines Clusters mit dem geringsten Abstand zugeordnet.
4. Auf Basis der entstandenen Cluster wird der Mittelpunkt eines jeden Clusters neu berechnet.
5. Wurde das vorgegebene Konvergenzkriterium noch nicht erreicht, wird der Algorithmus ab Schritt 2 erneut ausgeführt.

In der vorliegenden Arbeit wurde Clustering auf heterogene Variablen angewendet, dies bedeutet sowohl auf Variablen numerischen als auch kategorialen Typs. Um optimale Cluster zu erzielen, wurde die Summe der quadrierten Korrelationen als Distanz-Maß zum Mittelpunkt des jeweiligen Clusters maximiert, wobei die Korrelationen für numerische und kategoriale Variablen unterschiedlich berechnet wurde (siehe Formel 1.8) (Chavent et al., 2012).

Die Berechnung des Mittelpunkts eines Clusters k wird wie in Formel 1.8 definiert bestimmt (Chavent et al., 2012).

$$y_k = arg \max_{u \in R^n} \{ \sum_{x_j \in C_k} r^2_{u,x_j} + \sum_{z_j \in C_k} \eta^2_{u|z_j} \} \tag{1.8}$$

wobei r^2 die quadrierte Pearson-Korrelation zwischen numerischen Variablen und η^2 das Korrelations-Verhältnis zwischen kategorialen Variablen beschreibt (siehe Formel 1.10) (Chavent et al., 2012).

Daraus ergibt sich sie Homogenität eines Clusters wie folgt:

$$H(C_k) = \sum_{x_j \in C_k} r^2_{x_j,y_k} + \sum_{z_j \in C_k} \eta^2_{y_k|z_j} \tag{1.9}$$

wobei y_k der Cluster Mittelpunkt (siehe Formel 1.8) ist, r^2 die quadrierte Pearson-Korrelation zwischen numerischen Variablen und dem Cluster Mittelpunkt sowie η^2 die Korrelation zwischen kategorialen Variablen und dem Cluster Mittelpunkt beschreibt (siehe Formel 1.10) (Chavent et al.,

2012).

Das Korrelations-Verhältnis zwischen kategorialen Variablen wurde dabei wie in Gleichung 1.10 formalisiert.

$$\eta^2_{u|z_j} = \frac{\sum_{s \in M_j} n_s(\bar{u_s} - \bar{u})^2}{\sum_{i=1}^{n}(u_i - \bar{u})^2} \tag{1.10}$$

wobei n_s die Auftretens-Häufigkeit von Kategorie s ist, $\bar{u_s}$ der Mittelwert von u berechnet auf Basis aller Datenpunkte, bei denen Kategorie s beobachtet wurde und $\bar{u}$ der Mittelwert von u ist (Chavent et al., 2012).
Der verwendete Algorithmus bricht ab, wenn sich nach Berechnung der neuen Gruppenmittelpunkte die Zuordnung der Datenpunkte zu den Clustern nicht mehr verändert hat (Chavent et al., 2012).

1.4 Empirische Studien

Basierend auf den vorher beschriebenen Überlegungen, befassen sich die beiden Studien dieser Arbeit damit, den Einsatz von ML Algorithmen in der Fragebogenentwicklung zu explorieren.

In Kapitel 2 wird hierzu eine Studie beschrieben, in welcher Persönlichkeitseigenschaften auf Basis von Informationen vorhergesagt werden, welche bei der Erhebung eines bildbasierten Fragebogens gesammelt werden können. Anschließend wird überprüft, inwiefern diese Vorhersagen verwendet werden können, um Messmodelle für die Persönlichkeitsmessung zu erstellen.

In Kapitel 3 wird eine zweite Studie beschrieben, in welcher Persönlichkeitseigenschaften auf Basis von Panel-Daten vorhergesagt werden. Hierdurch wird der Forschungsfrage nachgegangen, inwiefern die Nutzung von ML Algorithmen und vorhandene Datensätze verwendet werden können, um einen Beitrag zur Validierung eines Fragebogens zu leisten.

Abschließend werden in Kapitel 4 die Ergebnisse der beiden Studien in Bezug auf die übergeordnete Fragestellung diskutiert, welche Potenziale und Herausforderungen im Einsatz von ML Methoden im Bereich der Fragebogenentwicklung bestehen.

Kapitel 2

Studie I: ML zur Ableitung von Messmodellen

2.1 Theoretischer Hintergrund

Persönlichkeitsmessung durch digitale Fußabdrücke Die enormen Entwicklungen der Digitalisierung bieten neben den in Kapitel 1.1.1 beschriebenen Methoden weitere Möglichkeiten zur Persönlichkeitserfassung mit Hilfe impliziter Methoden. Beispielsweise können die digitalen Fußabdrücke, die täglich hinterlassen werden, als implizites Maß für Persönlichkeitseigenschaften genutzt werden (Azucar, Marengo & Settanni, 2018). Dieses Forschungsgebiet wächst stetig. So werden beispielsweise in vielen Studien Informationen aus sozialen Netzwerken genutzt, um die Persönlichkeit von Nutzern vorherzusagen (Azucar et al., 2018). Als Prädiktoren werden hierbei vor allem die folgenden Informationen genutzt: veröffentlichte Texte (Ahmad & Siddique, 2017; Arnoux et al., 2017; Celli, 2012; Peng et al., 2015; Pratama & Sarno, 2015; Schwartz et al., 2013), Nutzerverhalten (Bai, Zhu & Cheng, 2012; Eftekhar, Fullwood & Morris, 2014; Golbeck et al., 2011; Kosinski et al., 2013; Quercia et al., 2011; Youyou, Kosinski & Stillwell, 2015) und (Profil-)Bilder (Eftekhar et al., 2014; Ferwerda, Schedl & Tkalcic, 2015; Guntuku, Qiu, Roy, Lin & Jakhetiya, 2015; Liu, Preotiuc-Pietro, Samani, Moghaddam & Ungar, 2016; Lovato et al., 2014; Segalin et al., 2017a; Wu, Chang & Yuan, 2015). Darüber hinaus konnten auch Zusammenhänge zwischen auf Basis der veröffentlichten Informationen in LinkedIn Profilen getroffenen Fremdeinschätzungen und selbstberichteten Persönlichkeitseigenschaften gezeigt werden (van de Ven, Bogaert, Serlie, Brandt & Denissen, 2017).

Für die vorliegende Studie sind vor allem die Ergebnisse in Bezug auf die Vorhersage von Persönlichkeitseigenschaften auf Grundlage von Fotos interessant. Fotos sind heutzutage ein wichtiges Kommunikationsmedium und Ausdruck der Identität

von Personen (van Dijck, 2008). Dementsprechend beinhalten Fotos relevante Informationen über die Persönlichkeit von Personen.

Im Forschungsbereich der automatisierten Persönlichkeitserkennung (Vinciarelli & Mohammadi, 2014) wurden bisher, wie oben beschrieben, vor allem Fotos betrachtet, welche von Personen in sozialen Netzwerken hochgeladen wurden (Eftekhar et al., 2014; Ferwerda et al., 2015; Guntuku et al., 2015; Liu et al., 2016; Segalin et al., 2017a; Wu et al., 2015). Ergebnisse dieser Studien zeigen beispielsweise, dass neurotische und extravertierte Personen mit Facebook-Erfahrung insgesamt mehr Fotos hochladen, gewissenhafte Personen mit Facebook-Erfahrung mehr Fotoalben selbst erstellen und mehr Videos hochladen, während hoch verträgliche Personen mehr Likes und Kommentare zu ihren Fotos bekamen (Eftekhar et al., 2014).

In einer Studie von Wu et al. (2015) wurden die Profilbilder von Facebook-Nutzern in verschiedene Kategorien eingeteilt (z. B. soziale Situationen, Familie, Sport). Anschließend wurden die Probanden gefragt, welche Persönlichkeitseigenschaften sie Personen zuschreiben würden, die als Profilfoto ein Bild aus einer solchen Kategorie verwenden. Die Autoren konnten einen Zusammenhang zwischen den selbstberichteten Persönlichkeitseigenschaften und den Einschätzungen Dritter anhand des Profilbildes zeigen (Wu et al., 2015). Darüber hinaus gaben die Probanden dieser Studie selbst an, dass ihre Profilbilder ihre Persönlichkeit widerspiegeln (Wu et al., 2015).

In einer anderen Studie zeigten Segalin, Perina, Cristani und Vinciarelli (2017b), dass Eigenschaften von Bildern, die zuvor als Favoriten gekennzeichnet wurden, Persönlichkeitseigenschaften vorhersagen können. Die Autoren konnten zeigen, dass die durch einen Dritten attribuierten Persönlichkeitseigenschaften besser vorhergesagt werden konnten als die selbstberichtete Persönlichkeitseigenschaftsausprägung. Darüber hinaus zeigten Marengo, Giannotta und Settanni (2017), dass die Identifikation mit verschiedenen Emojis einen Zusammenhang zu den Persönlichkeitseigenschaften Extraversion, Verträglichkeit und Neurotizismus aufweist.

Zusammenfassend zeigen die beschriebenen Studien, dass Fotos bei der Vorhersage von Persönlichkeitseigenschaften auf Basis digitaler Fußabdrücke als implizites Maß eine wichtige Rolle spielen.

Persönlichkeitsmessung durch bildbasierte Fragebögen Die Rolle von Bildern wurde auch in der expliziten Persönlichkeitserfassung erkannt. Bereits in den 1990er Jahren wurde ein Fragebogen zur bildbasierten Erfassung von Persönlichkeitseigenschaften bei Erwachsenen entwickelt (Paunonen, Jackson & Keinonen,

1990). Dieser Fragebogen nutzt Zeichnungen von Strichmännchen und legt das „System of Needs“ als Persönlichkeitskonzept zugrunde. Obwohl der Fragebogen in fünf Kulturkreisen validiert wurde (Paunonen, Zeidner, Engvik, Oosterveld & Maliphant, 2000) und später durch einen neuen Fragebogen eine Adaption an das Persönlichkeitskonzept der Big Five durchgeführt wurde (Paunonen, Ashton & Jackson, 2001), setzte sich dieser Fragebogen nicht als Standardinstrument der Persönlichkeitsmessung durch.

Im letzten Jahrzehnt wurden bildbasierte Fragebögen vor allem zur Erfassung verschiedener Konstrukte bei Kindern entwickelt (Döring, Blauensteiner, Aryus, Drögekamp & Bilsky, 2010; Mackiewicz & Cieciuch, 2016). Für diese Zielgruppe gibt es beispielsweise einen bildbasierten Fragebogen zur Erfassung von Persönlichkeitseigenschaften (Mackiewicz & Cieciuch, 2016) und einen zur Erfassung von Werten (Döring et al., 2010). Diese Fragebögen haben gemeinsam, dass sie an Kinderbücher angelehnte Zeichnungen nutzen, um die in Worten beschriebenen Verhaltensweisen zu illustrieren (Döring et al., 2010; Mackiewicz & Cieciuch, 2016).

In den genannten Fragebögen wurden einfache Zeichnungen als Stimulusmaterial genutzt, welche nur eine reduzierte Anzahl an Reizen beinhalten. Im Gegensatz hierzu scheint gerade die Vielzahl verschiedener Reize in Bildern interessant, wie wir ihnen im täglichen Leben immer wieder begegnen. Dafür spricht, dass Persönlichkeitseigenschaften neben inhaltlichen und ästhetischen Reizen in Bildern durch eine Vielzahl weiterer Reize angesprochen werden (Van Der Heide, D'Angelo & Schumaker, 2012) und Personen implizite Informationen aus Bildern ableiten, wie beispielsweise Werte und Einstellungen (Cristani, Vinciarelli, Segalin & Perina, 2015). Die genannten Studien zur Vorhersage von Persönlichkeitseigenschaften anhand normaler Fotos aus sozialen Netzwerken zeigen, dass auch komplexe Bilder Informationen über Persönlichkeitseigenschaften beinhalten. Daher wird in der vorliegenden Studie explorativ untersucht, inwiefern Bilder, die in Form eines Fragebogens dargeboten werden, Persönlichkeitseigenschaften von Personen vorhersagen können. Die Form der Bild-Präsentation in einem Fragebogen wird dabei bewusst gewählt. Hierdurch soll sichergestellt werden, dass die Personen sich dessen bewusst sind, dass sie Gegenstand einer Erhebung und Bewertung sind.

Ableitung der Forschungsfrage Die beschriebene Literatur zeigt einerseits, dass Fotos für die Vorhersage von Persönlichkeitseigenschaften relevant sind. Andererseits zeigt sie, dass es möglich ist, Persönlichkeitseigenschaften über bildbasierte Fragebögen zu erfassen. In der vorliegenden Studie werden diese beiden Erkenntnisse

kombiniert, sodass die Entwicklung eines Fragebogens auf Basis von Fotos exploriert wird. Diese inhaltliche Ausrichtung dient als Beispiel dafür, zu explorieren, inwiefern ML in der Fragebogenentwicklung verwendet werden kann. Die Verwendung von ML zusammen mit Kreuzvalidierungsverfahren hat gegenüber traditionellen Methoden den Vorteil, dass die Modelle daran evaluiert werden, wie gut sie Vorhersagen in neuen, zum Zeitpunkt der Modellerstellung unbekannten, Daten machen können (Breiman, 2001b; Shmueli, 2010). Darüber hinaus können durch ML neben linearen Zusammenhängen auch non-lineare Zusammenhänge modelliert werden (z. B. Hastie et al., 2009; James et al., 2013). Gerade bei der Exploration von neuem Stimulusmaterial scheint es sinnvoll zu explorieren, welche Art von Zusammenhang vorliegt.

Um alle Informationen, die bei der Beantwortung eines bildbasierten Fragebogens anfallen, für die Vorhersage von Persönlichkeitseigenschaften zu nutzen, werden in der Analyse neben den Bewertungen, wie sehr ein Bild zur Person passt, auch die Beantwortungszeiten pro Bild als Prädiktoren integriert. Darüber hinaus wird jeweils ein Interaktionsterm zwischen der Bewertung eines Bildes und der entsprechenden Beantwortungszeit gebildet und als Prädiktor integriert.

Diese Idee entspricht einerseits dem ML Ansatz mögliche Prädiktoren nicht vorab auszuschließen (Cheung & Jak, 2016; James et al., 2013), sondern den Algorithmus entscheiden zu lassen, welche Informationen zur Vorhersage wichtig sind. Auch Farnadi et al. (2016) haben gezeigt, dass sich die Vorhersagegüte nicht verbessert, wenn zunächst eine Variablenselektion anhand von Korrelationen mit dem Kriterium durchgeführt wird. Andererseits greift die Idee, Maße zur Beantwortungszeit zu integrieren, Forschung zur Antwortlatenz in der Persönlichkeitspsychologie auf (z. B. Ferrando & Lorenzo-Seva, 2007; Ranger & Ortner, 2011).

In diesem Forschungsbereich wird angenommen, dass schnelle Antworten auf Items bedeuten, dass in dem beantworteten Item eine typische bzw. für den Probanden wichtige Eigenschaft beschrieben ist, während langsame Antworten bedeuten, dass die beschriebenen Eigenschaften lediglich lose mit dem Selbst-Konzept der Person verbunden sind (Kuiper, 1981; Markus, 1977). Diese Eigenschaft von Antwortzeiten kann für dichotome Items in psychometrischen Modellen modelliert werden (Ferrando & Lorenzo-Seva, 2007; Ranger & Ortner, 2011). Da in der vorliegenden Studie allerdings ein intervallskaliertes Antwortformat verwendet wird, wird der beschriebene Zusammenhang zwischen ausgewählter Itemantwort und Beantwortungszeit durch einen Interaktionsterm modelliert. Diese Operationalisierung entspricht der Argumentation von Fekken und Holden (1992), dass zur Persönlichkeit der Person kongruente Items sowohl bei Zustimmung als auch bei Ablehnung

schneller verarbeitet werden.

Basierend auf den beschriebenen Überlegungen wird in der vorliegenden Studie die Forschungsfrage untersucht, inwiefern sich Persönlichkeitseigenschaften basierend auf den zuvor beschriebenen Variablen vorhersagen lassen. Dabei wird zudem der Frage nachgegangen, ob eine lineare oder non-lineare Modellierung den Zusammenhang zwischen Prädiktoren und Kriterium besser abbilden kann. Dies stellt in der Fragebogenentwicklung lediglich einen ersten Schritt dar. Neben dieser Evaluation, inwiefern die verwendeten Daten den gewünschten Messgegenstand abbilden können, ist es im Rahmen der Fragebogenentwicklung essentiell, ein Messmodell zu erstellen. Daher wird in der vorliegenden Studie zusätzlich exploriert, inwiefern die Ergebnisse der Vorhersagen dazu dienen können, psychometrische Messmodelle abzuleiten.

2.2 Methoden

Für die vorliegende Studie wurde ein bestehender Datensatz verwendet, welcher im Rahmen einer Abschlussarbeit erhoben wurde. Die Prä-Registrierung der Analyse sowie das präregistrierte und das finale R-Skript können im OSF abgerufen werden (https://osf.io/asdcn/).

2.2.1 Materialien, Vorgehen und Stichprobe

Materialien

Den Probanden wurden 30 Bilder gezeigt[1]. Um möglichst unterschiedliche Stimulusmaterialien zu erhalten, wurden für die Bildinhalte die Kategorien Landschaften, soziale Interaktionen, Stillleben und Tiere festgelegt. Als Grundlage für die Recherche lizenzfreier Bilder wurden Listen mit Adjektiven aus dem Manual des NEO-PI-R (Ostendorf & Angleitner, 2004), sowie aus den Studien von Goldberg (Goldberg, 1990; Goldberg, 1992) verwendet. Somit sollten die Bilder ähnlich zu dem Stimulusmaterial in der Studie von Van Der Heide et al. (2012) wichtige Aspekte der Persönlichkeitseigenschaften widerspiegeln. So wurden für Extraversion beispielsweise Bilder ausgewählt, die Personen umgeben von Freunden zeigten (Van Der Heide et al., 2012).

[1] Die verwendeten Bilder sind Abbildungen A.4 bis A.8 (Anhang A) zu entnehmen.

Da es sich um eine erste Exploration der Fragestellung handelt, wurden lediglich Bilder zu den drei Persönlichkeitseigenschaften Extraversion, Offenheit und Gewissenhaftigkeit integriert. Der Fragebogen enthielt am Ende 10 Bilder für jede Persönlichkeitseigenschaft und wurde online erhoben.

Zur Erfassung der Persönlichkeitseigenschaften wurde ein in der deutschsprachigen Sozialforschung etablierter Kurzfragebogen mit 15 Items (BFI-S) verwendet (Gerlitz & Schupp, 2005).

Vorgehen

Auf einer Einleitungsseite der Studie wurden das Ziel, die Teilnahmebedingungen, der Ablauf und der Umgang mit den Daten beschrieben. Durch das Klicken auf „Weiter“ erklärten sich die Teilnehmer mit den Teilnahmebedingungen einverstanden. Daraufhin wurden den Versuchspersonen nacheinander ein Bild sowie die Aussage „Dieses Bild passt zu mir.“ präsentiert, welche anhand einer 5-stufigen Skala[2] bewertet wurden. Da durch die Wahl der Bilder aus den beschriebenen Kategorien in einem Testlauf vor allem die Tierbilder von den Probanden als schwierig zu beantworten beschrieben wurden, wurden den Probanden zunächst 5 Bilder randomisiert gezeigt, die durch die Versuchsleiter als „einfach“ zu beantworten eingestuft wurden. Danach wurden nahtlos alle weiteren 25 Bilder randomisiert dargeboten.

Im Anschluss beantworteten die Probanden den BFI-S, einen Kurzfragebogen zur Erfassung von Persönlichkeitseigenschaften (Gerlitz & Schupp, 2005). Hier beantworteten die Probanden die Fragen auf einer 7-stufigen Skala, welche lediglich an den beiden Endpunkten verbal verankert war, während die Abstufungen nur durch eine Zahl verankert waren[3]. Die Items wurden alle auf einer Seite präsentiert, wobei die Reihenfolge der Items randomisiert war[4].
Die Beantwortung der 30 bildbasierten Fragen sowie der 15 BFI-S Items war obligatorisch.

Abschließend wurden die Probanden gebeten, demographische Angaben zu machen. Um die Anonymität der Daten in jedem Fall gewährleisten zu können, wurde

[2] Verankerungen 1= „sehr schlecht“, 2= „schlecht“, 3= „teils/teils“, 4= „gut“ und 5= „sehr gut“

[3] 1= „trifft überhaupt nicht zu“, 2, 3, 4, 5, 6, 7=„trifft voll zu“

[4] Um die Darstellung so nah wie möglich am Originalfragebogen realisieren zu können, wurden die Versuchspersonen vor Beantwortung des Persönlichkeitsfragebogens gefragt, ob sie die Befragung an einem mobilen Endgerät oder einem Computer durchführen. Anschließend wurde den Probanden eine für das entsprechende Endgeräte visuell optimierte Version des Fragebogens gezeigt.

Tabelle 2.1

Häufigkeitsverteilung Alterskategorien

Alterskategorie	*N*	Prozent
18-25	236	45
26-35	101	19
36-45	25	05
46-55	60	11
56-65	97	18
66+	10	2

Alter in den Kategorien „18-25 Jahre“, „26-35 Jahre“, „36-45 Jahre“, „46-55 Jahre“, „56-65 Jahre“, „66 Jahre und älter“ erfasst. Das Geschlecht wurde in 3 Kategorien erfasst („weiblich“, „männlich“, „sonstiges“).

Am Ende des Fragebogens wurden alle Bildquellen aufgelistet und den Versuchspersonen für ihre Teilnahme gedankt.

Stichprobe

Insgesamt füllten 530 Personen freiwillig den Fragebogen bis einschließlich aller BFI-S Items aus. Dies war in den Teilnahmebedingungen als Kriterium definiert, um die Daten für die Studie verwenden zu dürfen. Die Stichprobe bestand aus 66,4 % (351) weiblichen, 33,3 % (176) männlichen sowie 0,02 % (2) Teilnehmern mit sonstigem Geschlecht. Die Verteilung der Teilnehmer auf die verschiedenen Alterskategorien kann Tabelle 2.1 entnommen werden. Der Modalwert für das Alter lag in der Altersklasse 18-25 Jahre.

2.2.2 Auswertung

Im Rahmen der Datenvorverarbeitung wurden Werte als Ausreißer definiert, bei denen die Person mehr als 120 Sekunden zur Beantwortung eines Bildes benötigt hat. Bei diesen Werten ist davon auszugehen, dass die Probanden bei der Bearbeitung der Bilder abgelenkt waren und die Bearbeitung des Fragebogens somit zwischendurch unterbrochen haben. Bearbeitungszeiten > 120 Sekunden wurden durch fehlende Werte ersetzt. Da davon ausgegangen wird, dass die fehlenden Werte zufällig aufgetreten sind, wurden diese abschließend durch den Median der Beantwortungszeit

der jeweiligen Person imputiert. Anschließend wurden die Beantwortungszeiten innerhalb einer Person Median-zentriert (pro Bild wurde die absolute Abweichung vom Median berechnet), um personenbezogene Einflüsse auf die Beantwortungszeit herauszurechnen (Fekken & Holden, 1992)[5].

Den Hauptanalysen vorausgehend wurden deskriptive Analysen zu allen enthaltenen Roh-Variablen und korrelative Analysen zwischen den enthaltenen Variablen durchgeführt. Um eine Alpha-Fehler Inflation zu vermeiden, wurden die korrelativen Analysen Bonferroni-korrigiert.

In der vorliegenden Studie wurden die Kriterien in der Hauptanalyse als kontinuierliche Variablen realisiert, sodass es sich um Regressionsfragestellungen handelte.

Alle Auswertungen wurden mit der Statistik Software R (R Core Team, 2018) und unter Verwendung der folgenden Pakete durchgeführt: data.table (Dowle & Srinivasan, 2018), ggplot2 (Wickham, 2016), glmnet (Friedman, Hastie & Tibshirani, 2010), iml (Molnar, 2018), kernlab (Karatzoglou et al., 2004), mlr (Bischl et al., 2016), parallel (R Core Team, 2018), parallelMap (Bischl & Lang, 2015), plyr (Wickham, 2011), psych (Revelle, 2018), ranger (Wright & Ziegler, 2017), rpart (Therneau & Atkinson, 2018a), semPlot (Epskamp, 2019), stats (R Core Team, 2018) und xtable (Dahl, Scott, Roosen, Magnusson & Swinton, 2018).

Vorhersage von Persönlichkeitseigenschaften auf Basis von Bewertungen

In der Hauptanalyse der Studie wurden Benchmark-Experimente durchgeführt. Ziel der Benchmark-Experimente war einerseits, zu überprüfen, ob ein Zusammenhang zwischen den Prädiktoren und der als Kriterium definierten Persönlichkeitseigenschaft vorliegt. Andererseits sollte überprüft werden, welche Modellierung den Zusammenhang am besten annähern kann. Dazu wurde die Vorhersagegenauigkeit verschiedener Modellierungen zwischen Prädiktoren und Kriterium miteinander verglichen. Da Farnadi et al. (2016) zeigen konnten, dass bei der Vorhersage von Persönlichkeitseigenschaften keine signifikanten Unterschiede in der Vorhersageleistung zwischen univariaten und multivariaten Methoden besteht, wurde in der vorliegenden Studie ein univariater Ansatz verfolgt. Dies bedeutet, dass pro Vorhersage lediglich ein Kriterium betrachtet wird.

Als Kriterium der Vorhersagegüte wurden verschiedene Maße betrachtet: das R^2_C, die Standardabweichung des $R^2_C s$ $SD(R^2_C)$, der mittlere quadrierte Vorhersagefehler (MSE), Spearman-Rho (ρ) und die Rechenzeit (t). Eine Beschreibung dieser Maße

[5] Die in diesem Abschnitt beschriebenen Vorverarbeitungsschritte wurden nach der Präregistrierung des Analyse-Skripts aufgenommen.

ist Kapitel 1.3.1 zu entnehmen.

In bisherigen Studien wurden die folgenden Methoden verwendet, um Persönlichkeit vorherzusagen: SVM (Farnadi et al., 2016), Entscheidungsbaum (Farnadi et al., 2016), LASSO (Cristani et al., 2015; Guntuku et al., 2015; Lovato et al., 2014; Youyou et al., 2015), lineares Modell (Segalin et al., 2017a) und RF (Farnadi et al., 2016). Daher wurden in der vorliegenden Studie diese Methoden in einem Benchmark Experiment verglichen. Die Vorhersageleistung dieser Modelle wurde mit der Vorhersage eines Baseline-Modells (Featureless) verglichen, welches immer den Mittelwert des Kriteriums vorhersagt.

Für die Learner SVM[6], Entscheidungsbaum[7], LASSO[8] und RF[9] wurden die optimalen Werte der variablen Parameter in einer 10-fachen inneren Kreuzvalidierung ermittelt. Hierfür wurden verschiedene Werte im Rahmen eines Grid-Tunings miteinander verglichen (Probst et al., 2018).

Im Rahmen der Benchmark Experimente wurden für fehlende Werte bei intervallskalieten Variablen der Median und bei kategorialen Variablen der Modus imputiert[10].

Um robuste Werte für die Vorhersagegenauigkeit zu erhalten, wurde eine 10-fach wiederholte 10-fache Kreuzvalidierung durchgeführt. Diese Strategie wurde verwendet, da sie sich in einer Simulationsstudie von Kim (2009) bei Stichprobengrößen über 140 Personen als stabil erwiesen hat.

---

[6] Tuning des Kosten-Werts C (von -10 bis 10 in Schritten der trafo-Funktion (x), 2^x), des *kernel* (mit Werten der Radialen Basis Funktion, Gaussian) und des Sigma (mit Werten von -10 bis 10 in Schritten der trafo-Funktion$(x), 2^x$)

[7] Tuning des Komplexitätsparameters *cp* (von 0.00 bis 0.05 in 0.01 Schritten)(Nach Durchführung der Analyse die Datenvorverarbeitung verändert wurde, sodass die Analysen erneut gerechnet wurden, wurden die Tuning-Ergebnisse der initialen Analyse evaluiert und der Bereich für das Tuning des Komplexitätsparameter verringert, um den computationalen Aufwand zu verringern. In der initialen Analyse erreichte der Entscheidungsbaum trotz eines aufwendigeren Tunings keine bessere Performanz als die anderen Algorithmen), der maximalen Tiefe eines jeden Knotens im finalen Baum *maxdepth* (Werte 5, 10, 15, 20, 30), der Anzahl der Beobachtungen in jedem Blatt *minbucket* (Werte 5, 10, 15, 20, 30, 40, 50) und der minimalen Anzahl an Beobachtungen, die in einem Knoten vorhanden sein muss, um einen Split zu erzeugen *minsplit* (Werte 5, 10, 15, 20, 30, 40, 50, 60)

[8] Tuning anhand der Grundeinstellungen des Pakets glmnet (Friedman et al., 2010) für den Befehl cv.glmnet

[9] Tuning der Parameter *mtry* (Werte 1, 2, 3, 4, 5, 6, 7, 8) und *min.node.size* (Werte 5,15,25,35,45,55,65)

[10] Diese Imputationsmethode weicht von der präregistrierten Methode ab und wurde gewählt, da sie ausreißerrestistent ist.

Prädiktoren Als Prädiktoren wurden in der vorliegenden Studie die Bewertung, inwiefern ein Bild zur Versuchsperson passt, die Beantwortungszeit sowie ein Interaktionsterm zwischen Bewertung des Bildes und entsprechender Beantwortungszeit verwendet.

Durch die Einführung eines Interaktionsterms bekommen Antworten, welche eine niedrige Bewertung und lange Antwortzeit haben, einen ähnlichen Wert wie Antworten, die eine hohe Bewertung und kurze Antwortzeit haben. Eine Bewertungszeit im mittleren Bereich spricht somit für eine mittlere Ausprägung der Eigenschaft. Bei einer mittleren Bewertung des Bildes kann die Beantwortungszeit demnach Informationen darüber liefern, wie sicher die gegebene Antwort ist.

Kriterien Es wurden fünf verschiedene Vorhersagen mit jeweils einem Kriterium gemacht. Das Kriterium war jeweils eine Persönlichkeitsdimension (Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit, Neurotizismus). Für jede Person wurde der Skalenmittelwert für jede Persönlichkeitsdimension berechnet[11].

Wenngleich im zugrundeliegenden Datensatz Bilder auf Basis der Beschreibung von drei Persönlichkeitseigenschaften ausgewählt wurden, wurden alle fünf Dimensionen als Kriterien betrachtet, da nicht ausgeschlossen werden konnte, dass die Auswahl der Bilder fehlerhaft war. Dieses Vorgehen schien im Sinne eines Multitrait-Multimethod Ansatzes (Campbell & Fiske, 1959) sinnvoll, weil dadurch getestet werden konnte, inwiefern in den Bildern wirklich für die ausgewählten Persönlichkeitseigenschaften relevante Stimuli enthalten sind.

Ableitung eines Messmodells aus den Vorhersagen [12]

Die zuvor beschriebenen Methoden dienen der Evaluation, inwiefern die verwendeten Prädiktoren zur Vorhersage von Persönlichkeitseigenschaften verwendet werden können. Um darüber hinaus zu evaluieren, inwiefern Messmodelle abgeleitet werden können, wurden auf Basis der Ergebnisse der Vorhersagen konfirmatorische Faktorenanalysen gerechnet. In diesen wurden die wichtigsten Prädiktoren aus den Vorhersagemodellen als manifeste Variablen für die jeweilige latente Variable modelliert.

Da die Bilder in der vorliegenden Studie theoriegeleitet ausgewählt wurden, wurde neben dem zuvor beschriebenen komplett datengetriebenen Ansatz zusätzlich ex-

[11] Obwohl bei der Erhebung der Datensätze aus Darstellungsgründen unterschieden wurde, ob die Probanden die Umfrage an einem mobilen Endgerät oder einem Computer durchgeführt haben, wurde nicht zwischen den verwendeten Endgeräten unterschieden, da in einer entsprechenden Vergleichsstudie gezeigt werden konnte, dass die Art des Endgeräts keinen Einfluss auf die Ergebnisse des Fragebogens hat (de Bruijne & Wijnant, 2013).

[12] Diese Analyse geht über die prä-registrierten Analysen hinaus.

ploriert, wie sich die konfirmatorische Faktorenanalyse verändert, wenn lediglich die Prädiktoren von Bildern im Modell enthalten sind, die zur Messung dieser Persönlichkeitseigenschaft intendiert waren.

Als Maße für die Passung des Modell-Fits werden in der vorliegenden Studie der χ^2-Test, der Comparative Fit Index (CFI), der Root Mean Squared Error of Approximation (RMSEA) und das Standardised Root Mean Square Residual (SRMR) betrachtet. Bei einem guten Modell sollten der χ^2-Test nicht signifikant werden, der CFI einen Wert > 0.95, der RMSEA einen Wert < 0.06 und der SRMR einen Wert < 0.08 aufweisen (Hu & Bentler, 1999). Der χ^2-Test wird bei großen Stichproben schnell abgelehnt, der CFI wird durch die Stichprobengröße nicht beeinflusst (Hooper, Coughlan & Mullen, 2008). Während der RMSEA sparsame Modelle bevorzugt, erhält der SRMR kleine Werte, wenn viele Parameter enthalten sind und die Stichprobe groß ist (Hooper et al., 2008).

2.3 Ergebnisse

Im Folgenden werden die Ergebnisse der Auswertung dargestellt.

2.3.1 Deskriptive Ergebnisse

Deskriptive Statistiken (Mittelwert, Standardabweichung, Median, Minimum, Maximum, Schiefe, Kurtosis, Standardfehler) der Bildbewertungen sind Tabelle A.1 (Anhang A) zu entnehmen. Diese zeigen, dass für alle Bilder alle Antwortkategorien belegt sind.

Darüber hinaus sind Tabelle A.2 (Anhang A) deskriptive Statistiken der Bearbeitungszeit zu entnehmen. Diese zeigen, dass die Hälfte der Bilder im Median genauso schnell beantwortet wurden, wie die Teilnehmer im intra-individuellen Median benötigt haben, um ein Bild zu beantworten. Da nicht davon ausgegangen werden kann, dass die Probanden systematische Ähnlichkeiten in Bezug auf ihre Persönlichkeitsausprägungen aufweisen, können die Abweichungen analog zu itembasierter Fragebogen-Forschung darauf hindeuten, dass Unterschiede in der Item-Charakteristik bestehen (Dunn, Lushene & O'Neil, 1972).

Tabelle A.3 (Anhang A) sind deskriptive Statistiken des BFI-S zu entnehmen. Diese sind mit den von Hahn, Gottschling und Spinath (2012) publizierten Werten vergleichbar.

Zur Berechnung der Korrelationen zwischen den inkludierten Variablen wurde wie folgt vorgegangen. Da lediglich zwei Personen als Angabe des Geschlechts die

Kategorie „sonstiges“ gewählt haben, wurden diese beiden Personen in der Korrelationsanalyse ausgeschlossen und die Variable Geschlecht dummy-kodiert (weiblich = 0, männlich = 1). In Bezug auf die Altersklassen wurde eine neue, intervallskalierte Variable gebildet. Hier wurden die Werte mit steigendem Alter in der Kategorie aufsteigend gewählt beginnend von 1 bis 6. Diese Transformation wurde vorgenommen, um die Analyse zu vereinfachen. Insgesamt wurden 93 Korrelationen signifikant, wobei keine Auffälligkeiten erkennbar sind. Die gesamte Korrelationstabelle ist im OSF unter dem o. g. Link zu finden.

2.3.2 Ergebnisse der Vorhersage von Persönlichkeitseigenschaften

Im Folgenden werden die Ergebnisse für die jeweiligen Kriterien einzeln dargestellt.

Offenheit

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Offenheit können Tabelle 2.2 entnommen werden. Die Ergebnisse zeigen, dass Offenheit am besten durch einen RF vorhergesagt werden kann ($R^2_C = 0.19$, $SD = 0.08$). Darauf folgen SVM ($R^2_C = 0.18$, $SD = 0.09$) und LASSO ($R^2_C = 0.17$, SD= 0.06).

Im Folgenden werden die beiden Modelle RF und LASSO weiter betrachtet, da der RF zwar eine bessere Vorhersage macht, das LASSO aber einfacher zu interpretieren ist. Die Ergebnisse der SVM werden nicht weiter betrachtet, da die SVM ebenso ein schwer zu interpretierendes Modell darstellt und der RF eine bessere Vorhersagegüte erreicht.

Tabelle 2.2

Ergebnisse Benchmark Experiment zur Vorhersage von Offenheit

Learner	R^2_C	$SD(R^2_C)$	t	ρ	*MSE*
Featureless	-.02	.03	0.00	NA	1.32
Lineares Modell	-.11	.39	0.25	.45	1.40
Support Vektor Maschine	.18	.09	145.12	.45	1.06
LASSO	.17	.06	2.06	.49	1.08
Entscheidungsbaum	.13	.11	756.69	.39	1.11
Random Forest	.19	.08	1052.01	.47	1.05

Anmerkungen. $SD(R^2_C)$ = Standardabweichung des R^2_C, t = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error.

Permutation Feature Importance Tabelle 2.3 zeigt die zehn Prädiktoren mit der höchsten Permutation Feature Importance eines RF zur Vorhersage von Offenheit. Vier Prädiktoren stellen die Bewertung von Bilden dar, welche Offenheit repräsentieren sollen (O5, O2, O3, O8). Drei weitere Prädiktoren repräsentieren jeweils einen Interaktionsterm zwischen Bildbewertung und Beantwortungszeit eines Bildes, welches Offenheit repräsentieren soll (IO3, IO5, IO10). Darüber hinaus sind zwei Prädiktoren enthalten, die einen Interaktionsterm für Gewissenhaftigkeitsbilder darstellen (IG3, IG10), sowie ein Prädiktor, der einen Interaktionsterm für ein Extraversionsbild (IE5) darstellt.

Tabelle 2.3

Prädiktoren mit der höchsten Permutation Feature Importance zur Vorhersage von Offenheit mit einem RF

	Variable	FI
1	O5	1.79
2	O2	1.59
3	O3	1.54
4	IO3	1.26
5	O8	1.21
6	IG3	1.21
7	IG10	1.17
8	IO5	1.15
9	IE5	1.15
10	IO10	1.15

Anmerkungen. FI = Permutation Feature Importance. O1-O10 Bilder zu Offenheit, IO1-IO10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Offenheit, IG1-IG10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Gewissenhaftigkeit, IE1-IE10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Extraversion. Es werden nur die 10 wichtigsten Variablen dargestellt, wobei die Werte über 10 Wiederholungen gemittelt wurden. Tuning-Ergebnisse: mtry=3, min.node.size=25

Regressionsgewichte im LASSO Tabelle 2.4 zeigt, dass zur Vorhersage von Offenheit mit einem LASSO lediglich die Bewertungen der Bilder O5, O3 und O2 in das Modell eingegangen sind.

Global Surrogate Model Fittet man einen Entscheidungsbaum mit einer Tiefe von zwei auf die Vorhersage des Modells des RFs, so erklärt dieses Modell das ursprüngliche Modell mit einem $R^2 = 0.26$. Abbildung 2.1 kann entnommen wer-

Tabelle 2.4

Regressionsgewichte der relevanten Prädiktoren zur Vorhersage von Offenheit mit einem LASSO

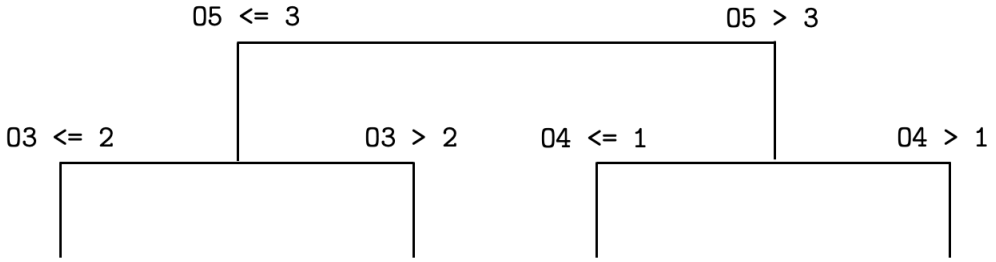
	Variable	*b*
	(Intercept)	3.834
1	O5	0.193
2	O3	0.126
3	O2	0.073

Anmerkungen. O1-O10 Bilder zu Offenheit. Im Modell wurden insgesamt 3 Prädiktoren beibehalten.

den, dass die Bewertungen der Bilder O5, O3 und O4 die wichtigsten Variablen zur Vorhersage von Offenheit darstellen.

Abbildung 2.1

Global Surrogate Model für Offenheit



Da das Ergebnis eines LASSO einfach zu interpretieren ist, wird hierfür kein Global Surrogate Model erstellt.

Gewissenhaftigkeit

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Gewissenhaftigkeit können Tabelle 2.5 entnommen werden. Die Ergebnisse zeigen, dass Gewissenhaftigkeit am besten durch ein LASSO vorhergesagt werden kann (R^2_C= 0.19, SD = 0.07).

Regressionsgewichte im LASSO Im Modell zur Vorhersage von Gewissenhaftigkeit mit einem LASSO sind zehn Prädiktoren enthalten geblieben. Diese können Tabelle 2.6 entnommen werden. Fünf der relevanten Prädiktoren stellen Bewertungen von Bildern zu Gewissenhaftigkeit dar (G3, G5, G1, G10, G4), darüber hinaus

Tabelle 2.5

Ergebnisse Benchmark Experiment zur Vorhersage von Gewissenhaftigkeit

Learner	R^2_C	$SD(R^2_C)$	t	ρ	MSE
Featureless	-.02	.03	0.00	NA	1.07
Lineares Modell	-.02	.23	0.11	.38	1.06
Support Vektor Maschine	.18	.10	145.75	.43	0.86
LASSO	.19	.07	1.37	.47	0.85
Entscheidungsbaum	.05	.11	765.29	.30	0.99
Random Forest	.13	.07	1027.54	.40	0.92

Anmerkungen. $SD(R^2_C)$ = Standardabweichung des R^2_C, t = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, MSE = Mean Squared Error.

sind Geschlecht und Alter der Probanden relevant sowie die Bewertung der Bilder E10, E5 und O7.

Tabelle 2.6

Regressionsgewichte der relevanten Prädiktoren zur Vorhersage von Gewissenhaftigkeit mit einem LASSO

	Variable	b
	(Intercept)	3.324
1	Geschlecht	-0.230
2	G3	0.206
3	G5	0.156
4	G1	0.088
5	Altersklasse 18-25	-0.087
6	G10	0.081
7	G4	0.036
8	E10	0.010
9	E5	0.010
10	O7	0.009

Anmerkungen. O1-O10 Bilder zu Offenheit, G1-G10 Bilder zu Gewissenhaftigkeit, E1-E10 Bilder zu Extraversion. weiblich = 0, männlich = 1. Im Modell wurden 10 Prädiktoren beibehalten.

Global Surrogate Model Da das Ergebnis eines LASSO einfach zu interpretieren ist, wird hierfür kein Global Surrogate Model erstellt.

Extraversion

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Extraversion können Tabelle 2.7 entnommen werden. Die Ergebnisse zeigen, dass Extraversion am besten durch eine SVM vorhergesagt werden kann ($R^2_C = 0.14$, $SD = 0.10$). Eine vergleichbare Vorhersagegüte wird durch ein LASSO ($R^2_C = 0.13$, $SD = 0.08$) und einen RF ($R^2_C = 0.13$, $SD = 0.07$) erreicht. Aufgrund der einfacheren Interpretierbarkeit und dem geringen Unterschied in der Vorhersagegüte wird neben der SVM im Folgenden auch das LASSO betrachtet.

Tabelle 2.7

Ergebnisse Benchmark Experiment zur Vorhersage von Extraversion

Learner	R^2_C	$SD(R^2_C)$	t	ρ	*MSE*
Featureless	-.02	.03	0.02	NA	1.80
Lineares Modell	-.08	.29	0.38	.39	1.87
Support Vektor Maschine	.14	.10	145.80	.39	1.51
LASSO	.13	.08	0.63	.39	1.53
Entscheidungsbaum	.08	.10	771.55	.34	1.62
Random Forest	.13	.07	1083.23	.39	1.53

Anmerkungen. $SD(R^2_C)$ = Standardabweichung des R^2_C, t = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error.

Permutation Feature Importance Tabelle 2.8 zeigt die Permutation Feature Importance der zehn wichtigsten Prädiktoren zur Vorhersage von Extraversion mit einer SVM. Hier sind vier Bewertungen von Extraversions-Bildern enthalten (E4, E2, E5, E3), Altersklasse, Geschlecht, ein Interaktionsterm für ein Extraversions-Bild (IE4), zwei Bewertungen von Gewissenhaftigkeits-Bildern (G5, G4) sowie eine Bewertung eines Offenheits-Bildes (O1).

Regressionsgewichte im LASSO Das Modell zur Vorhersage von Extraversion mit einem LASSO beinhaltet acht Prädiktoren. Vier dieser Prädiktoren stellen Bewertungen der Extraversions-Bilder dar (E4, E3, E8, E1), zwei stellen Bewertungen von Offenheits-Bildern dar (O4, O7), einer die Altersklasse 18-25 und ein letzter die Beantwortungszeit des Bildes G1. Die Regressionsgewichte können Tabelle 2.9 entnommen werden.

Tabelle 2.8

Prädiktoren mit der höchsten Permutation Feature Importance zur Vorhersage von Extraversion mit einer SVM

	Variable	FI
1	E4	1.29
2	E2	1.13
3	Altersklasse	1.10
4	G5	1.09
5	IE4	1.08
6	E5	1.07
7	Geschlecht	1.07
8	E3	1.07
9	G4	1.07
10	O1	1.07

Anmerkungen. FI = Permutation Feature Importance. O1-O10 Bilder zu Offenheit, IO1-IO10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Offenheit, IG1-IG10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Gewissenhaftigkeit, IE1-IE10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Extraversion. Es werden nur die 10 wichtigsten Variablen dargestellt, wobei die Werte über 10 Wiederholungen gemittelt wurden. Tuning-Ergebnisse: C=2.16, sigma=0.00456

Tabelle 2.9

Regressionsgewichte der relevanten Prädiktoren zur Vorhersage von Extraversion mit einem LASSO

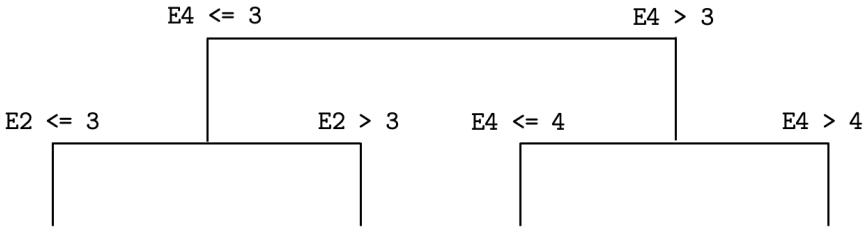
	Variable	*b*
	(Intercept)	2.289
1	E4	0.341
2	E2	0.145
3	E8	0.084
4	E1	0.050
5	O4	0.018
6	O7	0.014
7	Altersklasse 18-25	-0.008
8	ZG1	0.006
9	E7	0.003

Anmerkungen. O1-O10 Bilder zu Offenheit, E1-E10 Bilder zu Extraversion, ZG1-ZG10 Bearbeitungszeit der Bilder zu Gewissenhaftigkeit. Im Modell wurden 9 Prädiktoren beibehalten.

Global Surrogate Model Fittet man einen Entscheidungsbaum mit einer Tiefe von zwei auf die Vorhersage des Modells der SVM, so erklärt dieses Modell das ursprüngliche Modell mit einem $R^2 = 0.33$. Abbildung 2.2 kann entnommen werden, dass die Bewertungen der Bilder E4 und E2 die wichtigsten Variablen zur Vorhersage von Verträglichkeit darstellen.

Abbildung 2.2

Global Surrogate Model für Extraversion



Da das Ergebnis eines LASSO einfach zu interpretieren ist, wird hierfür kein Global Surrogate Model erstellt.

Verträglichkeit

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Verträglichkeit können Tabelle 2.10 entnommen werden. Die Ergebnisse zeigen, dass die Vorhersagegüte zur Vorhersage von Verträglichkeit deutlich geringer ist als zur Vorhersage der zuvor beschriebenen Persönlichkeitseigenschaften. Die beste Vorhersagegüte wird durch einen RF erreicht ($R^2_C = 0.02$, $SD = 0.06$), während Lineares Modell, LASSO und Entscheidungsbaum keine Vorhersage machen können und eine SVM ein R^2_C von 0.01 erreicht.

Permutation Feature Importance Tabelle 2.11 sind die zehn Variablen mit der höchsten Permutation Feature Importance im Modell des RF zur Vorhersage von Verträglichkeit zu entnehmen.

Global Surrogate Model Fittet man einen Entscheidungsbaum mit einer Tiefe von zwei auf die Vorhersage des Modells des RF, so erklärt dieses Modell das ursprüngliche Modell mit einem $R^2 = 0.12$. Abbildung 2.3 kann entnommen werden, dass die Bewertungen der Bilder E10 und E2 die wichtigsten Variablen zur Vorhersage von Verträglichkeit darstellen.

Tabelle 2.10

Ergebnisse Benchmark Experiment zur Vorhersage von Verträglichkeit

Learner	R^2_C	$SD(R^2_C)$	t	ρ	*MSE*
Featureless	-.02	.03	0.00	NA	1.00
Lineares Modell	-.30	.31	0.15	.18	1.26
Support Vektor Maschine	.01	.08	146.19	.23	0.97
LASSO	-.02	.03	0.55	NA	1.00
Entscheidungsbaum	-.04	.06	780.18	NA	1.02
Random Forest	.02	.06	1148.71	.21	0.96

Anmerkungen. $SD(R^2_C)$ = Standardabweichung des R^2_C, t = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error.

Tabelle 2.11

Permutation Feature Importance zur Vorhersage von Verträglichkeit mit einem RF

	Variable	FI
1	IE5	1.08
2	E2	1.07
3	IO4	1.07
4	IE2	1.06
5	ZG9	1.06
6	ZE7	1.06
7	ZO3	1.05
8	IE1	1.05
9	ZG4	1.05
10	ZE9	1.05

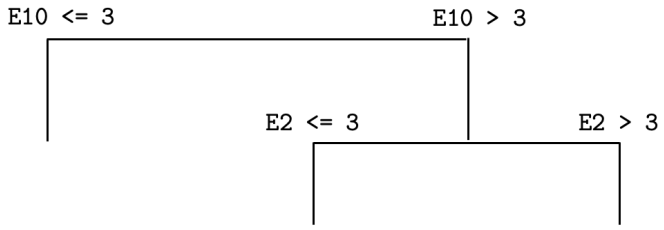
Anmerkungen. FI = Permutation Feature Importance. O1-O10 Bilder zu Offenheit, G1-G10 Bilder zu Gewissenhaftigkeit, E1-E10 Bilder zu Extraversion. ZO1-ZO10 Bearbeitungszeit der Bilder zu Offenheit, ZG1-ZG10 Bearbeitungszeit der Bilder zu Gewissenhaftigkeit, ZE1-ZE10 Bearbeitungszeit der Bilder zu Extraversion. IO1-IO10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Offenheit, IG1-IG10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Gewissenhaftigkeit, IE1-IE10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Extraversion. Es werden nur die 10 wichtigsten Variablen dargestellt, wobei die Werte über 10 Wiederholungen gemittelt wurden. Tuning-Ergebnisse: mtry=3, min.node.size=25

Neurotizismus

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Neurotizismus können Tabelle 2.12 entnommen werden. Die Ergebnisse zeigen, dass Neurotizismus am

Abbildung 2.3

Global Surrogate Model für Verträglichkeit



besten durch eine SVM vorhergesagt werden kann ($R^2_C = 0.09$, $SD = 0.11$).

Tabelle 2.12

Ergebnisse Benchmark Experiment zur Vorhersage von Neurotizismus

Learner	R^2_C	$SD(R^2_C)$	t	ρ	*MSE*
Featureless	-.02	.03	0.00	NA	1.74
Lineares Modell	-.22	.34	0.09	.30	2.05
Support Vektor Maschine	.09	.11	155.05	.33	1.54
LASSO	.07	.06	0.68	.33	1.58
Entscheidungsbaum	.02	.09	804.46	.23	1.66
Random Forest	.07	.06	1118.42	.32	1.57

Anmerkungen. $SD(R^2_C)$ = Standardabweichung des R^2_C, t = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error.

Permutation Feature Importance Tabelle 2.13 sind die zehn Variablen mit der höchsten Permutation Feature Importance im Modell des RF zur Vorhersage von Extraversion zu entnehmen.

Global Surrogate Model Fittet man einen Entscheidungsbaum mit einer Tiefe von zwei auf die Vorhersage des Modells der SVM, so erklärt dieses Modell das ursprüngliche Modell mit einem $R^2 = 0.35$. Abbildung 2.4 kann entnommen werden, dass die Bewertungen der Bilder E5, G5 und Geschlecht die wichtigsten Variablen zur Vorhersage von Offenheit darstellen.

Tabelle 2.13

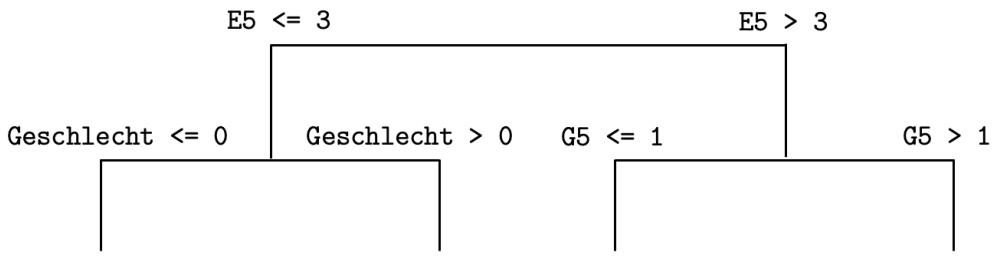
Permutation Feature Importance zur Vorhersage von Neurotizismus mit einer SVM

	Variable	FI
1	E5	1.11
2	O9	1.05
3	IG2	1.05
4	G7	1.05
5	ZE9	1.05
6	ZG7	1.05
7	IG10	1.04
8	G5	1.04
9	ZE5	1.04
10	ZE2	1.04

Anmerkungen. FI = Permutation Feature Importance. O1-O10 Bilder zu Offenheit, G1-G10 Bilder zu Gewissenhaftigkeit, E1-E10 Bilder zu Extraversion. ZO1-ZO10 Bearbeitungszeit der Bilder zu Offenheit, ZG1-ZG10 Bearbeitungszeit der Bilder zu Gewissenhaftigkeit, ZE1-ZE10 Bearbeitungszeit der Bilder zu Extraversion. IO1-IO10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Offenheit, IG1-IG10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Gewissenhaftigkeit, IE1-IE10 Interaktionen zwischen Bearbeitungszeit und Bildern zu Extraversion. Es werden nur die 10 wichtigsten Variablen dargestellt, wobei die Werte über 10 Wiederholungen gemittelt wurden. Tuning-Ergebnisse: C=0.463, sigma=0.00456

Abbildung 2.4

Global Surrogate Model für Neurotizismus



2.3.3 Ergebnisse der Ableitung von Messmodellen aus den Vorhersagemodellen

Um zu explorieren, inwiefern die Vorhersagemodelle verwendet werden können, um Messmodelle abzuleiten, wurden konfirmatorische Faktorenanalysen gerechnet. Da die Auswahl der Bilder in der vorliegenden Studie theoriegeleitet vorgenommen wurde, wurden aus den Prädiktoren, welche die höchste Vorhersagegüte erreichten, lediglich diejenigen als manifeste Variablen ausgewählt, die mit der entsprechenden Persönlichkeitseigenschaft zusammenhängen sollten. Für Verträglichkeit und Neurotizismus wurden dementsprechend keine Messmodelle erstellt.

Da die Zusammenhänge zwischen Prädiktoren und Kriterien univariat modelliert wurden, wurden für jede Persönlichkeitseigenschaft separate konfirmatorische Faktorenanalysen modelliert.

Offenheit

Basierend auf den Modellen der Vorhersage von Offenheit wurden zwei Messmodelle erstellt und im Rahmen einer konfirmatorischen Faktorenanalyse überprüft.

Als Modell basierend auf den Ergebnissen des RF wurde ein τ-kongenerisches Modell mit den Variablen O2, O3, O5 und O8 als manifeste Variablen spezifiziert. Dieses Modell kann basierend auf den Fit-Indizes beibehalten werden($\chi^2(2) = 3.116$, $p = .211$, p (Bollen-Stine Bootstrap) $= .227$, CFI $= 0.994$, RMSEA $= 0.032$, SRMR $= 0.022$).

Als Modell basierend auf den Ergebnissen des LASSO wurden die Prädiktoren, welche nicht auf Null gesetzt wurden (siehe Tabelle 2.4), als manifeste Variablen der latenten Variable Offenheit modelliert. Hierbei wurde ein essentiell τ-äquivalentes Modell spezifiziert, da ein τ-kongenerisches Modell nicht identifiziert war. Dieses Modell kann auf Basis der Fit-Indizes beibehalten werden ($\chi^2(2) = 2.192$, $p = .334$, p (Bollen-Stine Bootstrap) $= .368$, CFI $= 0.990$, RMSEA $= 0.013$, SRMR $= 0.026$).

Da für Offenheit zwei Modelle angenommen werden können, wurden dir komparativen Fit-Indizes AIC und BIC dieser Modelle verglichen. In beiden dieser Indizes erreichte das Modell basierend auf den Ergebnissen des LASSO geringere Werte (AIC $= 5124.172$, BIC $= 5141.263$) als das basierend auf den Ergebnissen des RF (AIC $= 6791.390$, BIC $= 6825.573$), weshalb ersteres bevorzugt wird.

Das Strukturgleichungsmodell kann Abbildung A.1 (Anhang A) entnommen werden.

Gewissenhaftigkeit

Basierend auf dem Ergebnis des LASSO wurde ein τ-kongenerisches Modell mit den Variablen G1, G3, G4, G5 und G8 spezifiziert. Dieses Modell kann aufgrund der Fit-Indizes beibehalten werden ($\chi^2(5) = 10.124$, $p = .072$, p (Bollen-Stine Bootstrap) = .082, CFI = 0.944, RMSEA = 0.044, SRMR = 0.030).

Das Strukturgleichungsmodell des Modells kann Abbildung A.2 (Anhang A) entnommen werden.

Extraversion

Basierend auf den Ergebnissen der SVM wurde ein τ-kongenerisches Modell erstellt, welches die Prädiktoren E2, E3, E4, E5 enthielt. Dieses Modell wurde aufgrund der Fit-Indizes abgelehnt, da es negative Schätzungen für Varianzen enthielt.

Basierend auf den im LASSO enthaltenen Prädiktoren wurde ein τ-kongenerisches Modell spezifiziert, welches die Variablen E1, E2, E4 und E8 enthielt. Dieses Modell kann aufgrund der Fit-Indizes beibehalten werden ($\chi^2(2) = 1.066$, $p = .587$, p (Bollen-Stine Bootstrap) = .579, CFI = 1.000, RMSEA = 0.000, SRMR = 0.014). Das Strukturgleichungsmodell dieser Modellierung kann Abbildung A.3 (Anhang A) entnommen werden.

2.4 Diskussion

In der vorliegenden Studie wurde explorativ untersucht, inwiefern durch die Daten, die bei der Beantwortung eines bildbasierten Fragebogens gesammelt werden, selbstberichtete Persönlichkeitseigenschaften vorhersagt werden können. Darüber hinaus wurde überprüft, inwiefern es möglich ist, psychometrische Messmodelle aus den Ergebnissen der Vorhersagen abzuleiten.

Hierzu wurden Probanden 30 Bilder präsentiert, die basierend auf Beschreibungen der Persönlichkeitseigenschaften Offenheit, Gewissenhaftigkeit und Extraversion ausgewählt wurden. Zur Vorhersage wurden sowohl die Selbstbewertung, inwiefern das jeweilige Bild zum Probanden passt, als auch die Bearbeitungszeit für jedes Bild sowie ein Interaktionsterm zwischen Bearbeitungszeit und Bewertung eines jeden Bildes als Prädiktoren integriert. Als Kriterium dienten in fünf verschiedenen Benchmark Experimenten jeweils der Mittelwert der Persönlichkeitseigenschaft, welcher anhand einer Big Five Kurzskala selbstberichtet wurde. Die Ergebnisse zeigen, dass sich selbstberichtete Persönlichkeitseigenschaften auf Basis der genannten Da-

ten vorhersagen lassen und die Modelle sich teilweise dazu eignen, Messmodelle abzuleiten.

Darüber hinaus zeigen die Ergebnisse, dass die Persönlichkeitseigenschaften, welche im Stimulusmaterial enthalten sein sollten, generell besser vorhergesagt werden konnten als die Persönlichkeitseigenschaften Verträglichkeit und Neurotizismus, die in den Bildern nicht repräsentiert waren. Dies deutet darauf hin, dass die Auswahl von Bildern auf Basis der in der Literatur verwendeten Adjektive vorteilhaft ist.

2.4.1 Diskussion der Studienergebnisse

Vorhersage von Persönlichkeitseigenschaften In den verschiedenen Benchmark Experimenten zeigten unterschiedliche ML Algorithmen die höchste Vorhersagegüte. Auffällig ist jedoch, dass bei den Persönlichkeitseigenschaften, welche das Stimulusmaterial abdecken sollte, die Modellierung mit einem LASSO eine ähnlich hohe Vorhersagegüte erreichte wie nicht-lineare Modellierungen durch SVM oder RF. Dies deutet darauf hin, dass eine lineare Modellierung der Zusammenhänge die zugrundelegenden Beziehungen gut abbilden kann. Dies ist als vorteilhaft einzuschätzen, da in der psychologischen Forschung traditionell von linearen Zusammenhängen ausgegangen wird (Nathans et al., 2012; Yin & Fan, 2001) und die Ergebnisse einer solchen Modellierung gut interpretierbar sind (James et al., 2013). Interpretierbarkeit scheint besonders in Bezug auf die Verwendung von ML zur Fragebogenentwicklung wichtig, da die Ableitung von wichtigen Variablen nur möglich ist, wenn das Modell gut verständlich ist.

Aus diesem Grund wurden in der vorliegenden Studie auf ML Algorithmen, welche die höchste Vorhersagegüte in einem Benchmark-Experiment erzielten und gleichzeitig als weniger gut interpretierbar gelten, Methoden zur Erhöhung dieser angewendet (Permutation Feature Importance und Global Surrogate Model)(Molnar, 2018). Diese Methoden lieferten unterschiedliche Ergebnisse in Bezug auf die Wichtigkeit einzelner Prädiktoren zur Vorhersage der Kriterien, sodass es schwierig ist, Informationen daraus abzuleiten. Darüber hinaus stellen die Ergebnisse der Methoden zur Erhöhung der Interpretierbarkeit lediglich eine Annäherung an das entsprechende Modell dar. Beispielsweise konnte in der vorliegenden Studie durch die jeweiligen Global Surrogate Modelle lediglich ein R^2 zwischen .12 und .35 des ursprünglichen Modells erklärt werden. In Zusammenhang mit der vergleichsweise geringen Vorhersagegüte der Modelle reduziert sich der Informationsgewinn daher noch einmal erheblich. Dies legt nahe, individuell zu entscheiden, inwiefern die Verwendung von Methoden zur Erhöhung der Interpretierbarkeit einen größeren Informationsgewinn

bringen als die Verwendung einfacherer zu interpretierender Algorithmen.

Darüber hinaus zeigen die verschiedenen Modellierungen für Offenheit und Extraversion, dass bei der Verwendung verschiedener Algorithmen, unterschiedliche Variablen für die Modelle wichtig sind. Der Argumentation von Yarkoni und Westfall (2017) folgend, sollte eine Interpretation daher mit einer skeptischen Haltung vorgenommen werden.

In Bezug auf die Frage, welche ML Algorithmen zur Vorhersage von Persönlichkeitseigenschaften verwendet werden sollten, legen die Ergebnisse der vorliegenden Studie eine Unterscheidung zwischen zwei Zielen nahe: der reinen Vorhersage und der Weiternutzung von Vorhersageergebnissen. Dadurch, dass nicht-lineare Modellierungen teilweise bessere Vorhersageergebnisse in Bezug auf die Kriterien erreichten, kann für das Ziel der reinen, möglichst präzisen Vorhersage empfohlen werden, verschiedene Algorithmen zu vergleichen, welche die Zusammenhänge sowohl linear als auch non-linear modellieren, um anschließend denjenigen zu verwenden, welcher die höchste Vorhersagegüte erreicht. Dies scheint gerade in der Anwendung ein gewünschtes Ziel (Shmueli, 2010).

Die psychologische Forschung befasst sich allerdings traditionell mit dem Verstehen von Verhalten und entwickelt basierend auf den Beobachtungen theoriebasierte Modelle, welche zur anschließenden Vorhersage von Verhalten verwendet werden können (z. B. Yarkoni & Westfall, 2017). Um ML Algorithmen für eine theoretische Weiterentwicklung zu nutzen, wie es beispielsweise auch Bleidorn und Hopwood (2019), Shmueli und Koppius (2011) und Shmueli (2010) vorschlagen, scheinen daher vor allem Algorithmen sinnvoll, welche gut interpretierbare Modelle erstellen. Nur bei dieser Art von Modellen scheint es möglich, wie von Shmueli (2010) empfohlen, relevante Prädiktoren zu identifizieren und somit induktiv Theorien weiterzuentwickeln (Shmueli, 2010; Shmueli & Koppius, 2011).

Ableitung von Messmodellen aus den Vorhersagemodellen Anhand der Vorhersagemodelle konnten für die drei Persönlichkeitseigenschaften Offenheit, Gewissenhaftigkeit und Extraversion Messmodelle mit Bildbewertungen als manifeste Variablen erstellt werden. Dabei zeigte sich, dass zur Weiterverwendung im Rahmen der Fragebogenentwicklung eine Integration linearer Modelle und bestehender Theorie am besten funktioniert. Dass die Ergebnisse der linearen Modelle eine bessere Basis für die Erstellung eines psychometrischen Messmodells darstellen, ist wenig überraschend, da eine konfirmatorische Faktorenanalyse ebenfalls lineare Zusammenhänge annimmt (Goretzko, Pham & Bühner, 2019).

Dennoch ist es erstaunlich, dass Modelle, welche lediglich Bilder enthalten, die sowohl theoriegeleitet als auch auf Basis der Ergebnisse des LASSO als Prädiktoren für die jeweilige Persönlichkeitseigenschaft identifiziert wurden, den besten Modellfit erreichen.

Der gute Modell-Fit kann darauf hindeuten, dass annehmbare Messmodelle erstellt wurden. Allerdings ist darauf hinzuweisen, dass die verwendeten Indizes und Cut-Off Werte irreführend sein können (Greiff & Heene, 2017). Auffällig ist beispielsweise, dass die Ladungen der einzelnen Bilder nicht besonders hoch sind und die Fehlerterme hohe Werte aufweisen. Darüber hinaus ist die Vorhersagegüte der zugrundeliegenden Modelle ebenfalls nicht besonders hoch, sodass kritisch hinterfragt werden muss, inwiefern die Messmodelle eine sinnvolle Messung der jeweiligen Persönlichkeitseigenschaft ermöglichen.

Darüber hinaus ist kritisch anzumerken, dass die Variablenselektion sowie die Überprüfung der Messmodelle in der vorliegenden Studie mit den gleichen Daten durchgeführt wurden (Loevinger, 1957; Smith & McCarthy, 1995; Tuckey, 1950).

In den Messmodellen, welche auf Basis der Ergebnisse der Vorhersagen erstellt wurden, werden die Persönlichkeitseigenschaften separat betrachtet. Die in diesen Messmodellen relevanten Variablen unterscheiden sich von den Variablen, welche unter Verwendung von explorativen Faktorenanalysen in den Modellen enthalten wären (Ergebnisse von entsprechenden EFAs können Tabellen A.4 bis A.6 in Anhang A entnommen werden). Daher sollte in zukünftigen Simulationsstudien überprüft werden, welche Methode besser in der Lage ist, vorliegende Strukturen zu erkennen.

Inkludierte Prädiktoren Die Bearbeitungszeit und der Interaktionsterm zwischen Bearbeitungszeit und Bildbewertung blieben in den Modellen des LASSO lediglich in einem Modell enthalten: Im Modell zur Vorhersage von Extraversion wurde die Bearbeitungszeit für das Bild G1 beibehalten. Da es hierfür keinerlei theoretische Begründung gibt, deutet dies darauf hin, dass die entsprechenden Variablen keinen zusätzlichen Informationsgewinn zur Vorhersage von Persönlichkeitseigenschaften bieten.

2.4.2 Stärken, Limitationen und Ausblick

Die vorliegende Studie exploriert als eine der ersten Studien den Einsatz von ML Algorithmen in der Fragebogenentwicklung und leistet somit einen wichtigen Beitrag dazu, mögliche Einsatz-Bereiche dieser Methoden in der psychologischen Forschung zu ermitteln.

Studiendesign Die Präsentation von Bildern an Stelle von stark vereinfachter Zeichnungen als Stimulusmaterial für einen Persönlichkeitsfragebogen stellt eine Stärke dar. Durch die direkte Darstellung von Situationen werden den Probanden alle Informationen direkt präsentiert (Wolfgang, 2005), sodass das Bild ein mentales Modell der dargestellten Situation hervorruft. Dies ist vorteilhaft, da hierdurch Variationen der mentalen Repräsentation verringert werden, welche im Rahmen des Übersetzungsprozesses von symbolbasierten Darstellungen (z. B. Text) auftreten können (Wolfgang, 2005). Dennoch bieten auch Bilder Interpretationsspielraum, da vor allem bei Bildern, welche mehrere Personen darstellen, nicht sicher ist, mit welcher sich die Probanden identifizieren und welche Eigenschaften sie dieser Person zuschreiben. Dies sollte in zukünftiger Forschung genauer untersucht werden.

Als Kriterium wurden in der vorliegenden Studie selbstberichtete Persönlichkeitseigenschaften aus einer Kurzskala verwendet. Da das Persönlichkeitskonzept der Big Five auf einem lexikalischen Ansatz beruht (McCrae & John, 1992; McCrae & Costa, 1987), stellt sich die generelle Frage, inwiefern dieses Konzept als Referenzrahmen für die Erfassung von Persönlichkeitseigenschaften ohne sprachliche Reize sinnvoll ist. Ein anderer Ansatz könnte sein, analog zum lexikalischen Ansatz einen großen Pool an persönlichkeitsrelevanten Bildern zu erstellen und anhand dessen ein bildbasiertes Persönlichkeitsmodell zu erstellen.

Zukünftige Forschung In der vorliegenden Studie war es möglich, Persönlichkeitseigenschaften vorherzusagen. Dennoch sollte explizit darauf hingewiesen werden, dass die Vorhersagen auf Basis der Modelle große Fehler haben. Dieser kann unter anderem dadurch entstanden sein, dass lediglich ein Teil der durch das Format vorhandenen Informationen verwendet wurde, da Bilder an sich bereits eine Fülle von Informationen beinhalten. Daher scheinen über die Bewertungen und Beantwortungszeiten hinaus auch Bildeigenschaften der Bilder, die als passend bewertet wurden, informativ für die Vorhersage der Persönlichkeitseigenschaften. Durch eine Integration solcher Informationen würden analog zur Studie von Wei et al. (2017) heterogene Informationen verwendet werden. Daher sollten auch die Bild-Eigenschaften, welche einerseits automatisch aus den Bildern extrahiert werden können, wie durchschnittliche Helligkeit, Sättigung und Farbpalette in zukünftigen Studien berücksichtigt werden, da diese Eigenschaften bereits als Prädiktoren von Persönlichkeitseigenschaften gezeigt werden konnten (z. B. Ferwerda et al., 2015; Liu et al., 2016; Segalin et al., 2017a). Aber auch die Integration inhaltlicher Eigenschaften scheint vielversprechend, da durch den Einsatz eines bildbasierten Fragebogens die Anzahl

der gezeigten Bilder begrenzt ist und es somit möglich ist, aus allen Bildern inhaltlich interpretierbare Eigenschaften abzuleiten (z. B. Ferwerda et al., 2015; Liu et al., 2016), welche ggf. über automatisch extrahierbare Eigenschaften hinaus gehen. Angelehnt an bisherige Studien scheinen vor allem die folgenden inhaltlichen Eigenschaften vielversprechend:

- Anzahl der Augenpaare, die in die Kamera gucken (Guntuku et al., 2015)
- Anzahl der Gesichter (z. B. Ferwerda et al., 2015; Penton-Voak, Pound, Little & Perett, 2006; Segalin et al., 2017a), die mindestens von der Seite zu sehen sind (Guntuku et al., 2015; Rule, Ambady & Adams, 2009)
- Kategorisierung der in Gesichtern ausgedrückten Emotionen (Walker & Vetter, 2016) bzw. Gesichtsausdrücken (Guntuku et al., 2015)
- Charakterisierung der Umgebung (öffentliches, privates Umfeld) (Guntuku et al., 2015)
- Anzahl der Personen männlichen/weiblichen Geschlechts (Guntuku et al., 2015)
- Gesamtzahl an Personen (Ferwerda et al., 2015)
- Anzahl lebloser Objekte (Lovato et al., 2014)
- Kategorisierung der Bildart (z. B. Fotomontage, Schnappschuss) (Guntuku et al., 2015)

Eine Möglichkeit, die Bildeigenschaften zu integrieren, wäre für jede Person die Bilder zu identifizieren, von denen sie angibt, dass diese sie besser als die Bilder im Mittel beschreiben. Anschließend könnte die Präferenz mit den Bildeigenschaften über eine Multiple-Intance-Regression (Ray & Page, 2001) modelliert werden. Diese Methode würde es ermöglichen, analog zur Studie von Segalin et al. (2017b) die Bildeigenschaften der Bilder, die eine Person besser als im Mittel beschreiben, zu verwenden, um die Ausprägung der Persönlichkeitseigenschaft dieser Person vorherzusagen.

Da angenommen wird, dass Persönlichkeitseigenschaften stabil sind (z. B. Costa & McCrae, 1988; McCrae & Costa, 1982), sollte außerdem überprüft werden, inwiefern die Bewertung von Bildern ebenfalls eine Stabilität aufweist. Hierzu könnte beispielsweise überprüft werden, ob die Test-Retest-Reliabilität hoch ist (Fiske & Rice, 1955). Andererseits könnte auch untersucht werden, inwiefern die Bewertung der Bilder von stark variablen Faktoren wie beispielsweise Stimmung, Situation, Befragungsumgebung abhängt, um den Trait- und State-Anteil in der Messung zu bestimmen (Steyer, Schmitt & Eid, 1999). Solche Studien könnten zusätzlich dabei

helfen, einzuschätzen, ob die Messung von Persönlichkeitseigenschaften mit Hilfe von Bildern überhaupt sinnvoll ist. Denn nur, wenn ein Großteil der Varianz zwischen verschiedenen Personen durch Persönlichkeitseigenschaften erklärbar sind, scheint dieses Vorhaben sinnvoll.

Um zu verstehen, welche Aspekte in einem Bild relevant sind, könnten zudem Studien mit Eye-Tracking durchgeführt werden. Dies könnte einerseits dazu beitragen, relevante Inhalte zur Vorhersage von Persönlichkeitseigenschaften zu identifizieren. Andererseits könnte die Integration von Eye-Tracking Daten auch die Vorhersage von Persönlichkeitseigenschaften verbessern, da es denkbar ist, dass Personen aufgrund unterschiedlicher Persönlichkeitseigenschaften verschiedene Bild-Elemente fixieren. Dass Traits beeinflussen, welche Bildelemente für wie lange angeschaut werden, konnte bereits in verschiedenen Studien gezeigt werden (z. B. Isaacowitz, 2005; Kaspar & König, 2011; Rauthmann, Seubert, Sachse & Furtner, 2012; Wilkowski, Robinson, Gordon & Troop-Gordon, 2007).

2.4.3 Fazit

Die Ergebnisse der vorliegenden Studie zeigen, dass in der Erfassung von Persönlichkeitseigenschaften durch einen bildbasierten Fragebogen grundsätzlich Potenzial steckt, da es möglich war, Persönlichkeitseigenschaften vorherzusagen. Aufgrund der beschriebenen Limitationen der vorliegenden Studie ist es allerdings zu früh, Schlüsse zu ziehen, wie genau Bilder zur Erfassung von Persönlichkeitseigenschaften aussehen sollten.

Darüber hinaus deuten die Ergebnisse darauf hin, dass ML in der Fragebogenentwicklung vor allem als Explorationsmedium gesehen werden kann, welches eingesetzt wird um herauszufinden, an welchen Stellen Zusammenhänge in zukünftigen Studien überprüft werden sollten. Inwiefern diese Methoden allerdings gegenüber Methoden, die traditionell in der Fragebogenentwicklung verwendet werden (CFA und EFA), bevorzugt werden können, muss in Simulationsstudien geprüft werden.

Kapitel 3

Studie II: Einsatz von ML Algorithmen zur Vorhersage von Fragebogenscores mit Panel-Daten - Ein Validierungsansatz für Fragebögen?

3.1 Einleitung

In der psychologischen Forschung stellt die Rekrutierung geeigneter Probanden eine zentrale Herausforderung dar. Durch Poweranalysen können optimale Stichprobengrößen auf Basis erwarteter Effektgrößen und einem festgelegten Alphafehlerniveau bestimmt werden (Tressoldi, 2012). Um Effekte, welche in der Psychologie mittlere Effektgrößen von $d = 0.50$ aufweisen (Bakker, van Dijk & Wicherts, 2012), durch hohe Power abzusichern und somit Fehlschlüsse zu vermeiden, müssen große Stichproben erhoben werden (Tressoldi, 2012). Daher stehen Forscher immer wieder vor der Herausforderung, große Stichproben zu erheben. Diese Problematik verstärkt sich in der Fragebogenentwicklung zusätzlich, da bis zur Fertigstellung eines Fragebogens mehrere Iterationen durchlaufen werden müssen (siehe Kapitel 1.2.1), welche jedes Mal neue, ausreichend große und möglichst repräsentative Stichproben erfordern (Clark & Watson, 1995; Loevinger, 1957; Smith & McCarthy, 1995; Tuckey, 1950). Dies ist besonders wichtig, da die Verwendung vieler kleiner Stichproben mit den dazugehörigen Tests die Wahrscheinlichkeit falsch-positiver Ergebnisse erhöht (Bakker et al., 2012).

Dass die Rekrutierung von Stichproben im psychologischen Umfeld nicht einfach ist, ist daran zu erkennen, dass die Stichproben häufig primär aus (Psychologie -)Studierenden bestehen (Arnett, 2008; Henrich, Heine & Norenzayan, 2010b; Peterson, 2001; Rad, Martingano & Ginges, 2018; Wintre, North & Sugar, 2007). Eine typische Stichprobe in der psychologischen Forschung stammt aus einer jungen Population mit einem Bildungsniveau aus westlichen Ländern, vor allem den USA (Arnett, 2008; Henrich et al., 2010b; Rad et al., 2018). Um die dadurch entstehenden, offensichtlichen Nachteile in Bezug auf die Generalisierbarkeit (Henrich, Heine & Norenzayan, 2010a, 2010b; Peterson, 2001; Rad et al., 2018; Wintre et al., 2007) zu umgehen, sollten repräsentativere Stichproben verwendet werden. Hierbei ist zu beachten, dass die geeignete Stichprobe im Rahmen der Fragebogenentwicklung von der Zielgruppe für diesen abhängt.

Rekrutierung repräsentativer Stichproben Da die Abhängigkeit von studentischen Stichproben in der Psychologie und die daraus resultierenden Nachteile bereits umfassend in der Literatur diskutiert wurden (Wintre et al., 2007), gibt es mittlerweile verschiedene Ansätze, um Probanden zu rekrutieren. Beispielsweise wurden Crowd-Sourcing Plattformen als Rekrutierungsort identifiziert (Behrend, Sharek, Meade & Wiebe, 2011; Buhrmester, Kwang & Gosling, 2011; Litman, Robinson & Abberbock, 2017; Steelman, Hammer & Limayem, 2014). Eine weitere Möglichkeit, Forschung mit repräsentativeren Stichproben durchzuführen, ist der Einsatz von Panels, welche vor allem im Bereich der Marktforschung beliebt sind (Steelman et al., 2014). Solche Panels existieren darüber hinaus auch für Forschungszwecke (z. B. GESIS Panel (GESIS, 2018; Shmueli, 2017)). Im Folgenden wird das GESIS Panel genauer beschrieben, da dies als Grundlage für die vorliegende Arbeit dient. Im Rahmen der Befragungen im GESIS Panel wird eine Vielzahl von Informationen über Personen erhoben und Forschern die Möglichkeit gegeben, eigene Fragebögen einzureichen, um Daten zu erheben. Die dadurch erhobenen Datensätze können von Forschern beantragt werden und enthalten nicht nur viele Variablen, sondern meist auch große, repräsentative Stichproben. Diese verringern zudem die Wahrscheinlichkeit für Stichprobenfehler (z. B. Kosinski, Wang, Lakkaraju & Leskovec, 2016).

Ableitung der Forschungsfrage Die beschriebene Literatur zeigt einerseits, dass gerade in der Fragebogenentwicklung mehrfach große Stichproben erhoben werden müssen. Andererseits zeigt sie, dass Forscher durch Angebote wie das GESIS Panel (GESIS, 2018) Zugang zu umfassenden Datensätzen mit großen, repräsentativen Stichproben haben können. Aus ökonomischen Überlegungen zur optimalen Nutzung von Ressourcen, scheint es daher sinnvoll, zu überprüfen, inwiefern die

vorhandenen Datensätze im Rahmen der Fragebogenentwicklung genutzt werden können.

In der vorliegenden Studie wird dazu die Evaluation der Validität (siehe Kapitel 1.2.1) genauer betrachtet und exploriert, inwiefern bereits vorhandene Daten aus Panel-Daten diesen Evaluationsschritt unterstützen können. Hierzu wird ein bereits vorhandener und im Rahmen der Befragung im GESIS Panel (GESIS, 2018) eingesetzter Persönlichkeitsfragebogen (BFI-10 (Rammstedt, Kemper, Klein, Beierlein & Kovaleva, 2014)) verwendet.

Klassischerweise werden zur Evaluation der Validität Korrelationsstudien durchgeführt (Bleidorn & Hopwood, 2019). Da bei der Verwendung von Panel-Daten allerdings eine so große Menge an Variablen zur Verfügung steht, wäre es enorm aufwendig, alle Korrelationen zu betrachten. Daher wird in der vorliegenden Studie zusätzlich exploriert, inwiefern ML Algorithmen in diesem Rahmen sinnvoll einzusetzen sind. Hierzu wird einerseits überprüft, inwiefern unterschiedliche vorgelagerte Algorithmen zur Dimensionsreduktion die Vorhersagegüte erhöhen können. Darüber hinaus werden unterschiedliche Algorithmen zur Modellierung zwischen den Prädiktoren und Kriterien verwendet. Mit Hilfe dieser soll exploriert werden, inwiefern die Zusammenhänge besser non-linear oder linear modelliert werden können.

3.2 Methode

3.2.1 Materialien und Stichprobe

Material

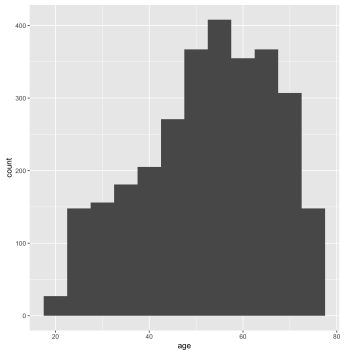
In allen Analysen dieser Studie wurden die Daten aus dem GESIS Panel Version ZA5665 verwendet (GESIS, 2018). Diese Daten zeichnen sich dadurch aus, dass eine für die deutsche Bevölkerung repräsentative Stichprobe in zweimonatigen Abständen zu verschiedenen Themengebieten befragt wird. Die Themen werden für jede Befragung neu festgesetzt. Die verwendeten Daten wurden zwischen 2016 und Dezember 2017 gesammelt (Befragungswellen *dc-ed*).

Stichprobe

Die finale Stichprobe bestand aus 2974 Personen, wovon 49,76% (1480) weiblich und 49,70% (1478) männlich waren. Die Stichprobe enthält alle Personen, die an den acht betrachteten Wellen teilgenommen haben. Die Altersverteilung der Probanden ist Abbildung 3.1 zu entnehmen.

Abbildung 3.1

Altersverteilung in der Stichprobe



3.2.2 Auswertung

Alle Auswertungsskripte sind dem OSF (https://osf.io/b9ndx) zu entnehmen.

Die Auswertungen wurden mit der Statistik Software R (R Core Team, 2018) und unter Verwendung der folgenden Pakete durchgeführt: ClustOfVar (Chavent et al., 2012), data.table (Dowle & Srinivasan, 2018), devtools (Wickham, Hester & Chang, 2019b), dpyr (Wickham, François, Henry & Müller, 2019a), glmnet (Friedman et al., 2010), haven (Wickham & Miller, 2019), liquidSVM (Steinwart & Thomann, 2017), mlr (Bischl et al., 2016), mlrCPO (Binder, 2019), parallelMap (Bischl & Lang, 2015), PCAmixdata (Chavent, Kuentz-Simonet, Labenne & Saracco, 2017), plyr (Wickham, 2011), psych (Revelle, 2018), ranger (Wright & Ziegler, 2017), readr (Wickham, Hester & François, 2018) und xtable (Dahl et al., 2018).

Vorhersage von Persönlichkeitseigenschaften auf Basis von Panel-Daten

Datenvorverarbeitung Um die Rohdaten für die Analysen nutzbar zu machen, wurden diese vorverarbeitet. Hierbei wurde sich in Teilen an dem Vorverarbeitungsskript von Pargent und Gönna (2018) orientiert[1].

Zunächst wurde festgelegt, dass die Variablen aus den Befragungswellen *dc* bis *ed* für die vorliegende Studie verwendet werden. Somit konnten zwei Kohorten an Teilnehmern der Panel Befragungen inkludiert werden, welche an acht Befragungszeitpunkten teilnahmen.

Aus den Berfragungswellen wurden die inhaltlichen Variablen als Prädiktoren eingeschlossen. Aus diesem Pool wurden alle Variablen ausgeschlossen, die eine Zuordnung zu Experimentalgruppen darstellen, Variablen, die durch experimentelle

[1] Das Vorverarbeitungsskript von Pargent und Gönna (2018) ist https://osf.io/zpse3/zu entnehmen

Variation nur einem Teil der Teilnehmer gezeigt wurden und Variablen, in denen Persönlichkeitseigenschaften abgefragt wurden (insgesamt 246 Variablen).

Nach dieser Selektion wurde über alle Variablen die mittlere Anzahl an fehlenden Werten, welche in den Panel-Rohdaten enthalten waren, sowie die dazugehörige Standardabweichung bestimmt. Anschließend wurden alle Variablen ausgeschlossen, welche mehr als eine Standardabweichung als die mittlere Anzahl an fehlenden Werten aufwiesen (62 Variablen).

Die resultierenden 1993 Variablen wurden weiter verarbeitet. Einerseits wurden alle Antwortformate dahingehend überprüft, ob eine numerische Kodierung (z.B. in Form von Ratingskalen) bzw. eine nominale Kodierung mit genau oder mehr als zwei Antwortkategorien vorlag.

Neben der eigentlichen Antwortskala waren bei vielen Variablen ein oder mehrere zusätzliche Antwortkategorien mit den Labels „weiß nicht“, „trifft nicht zu“, „Das möchte ich nicht beantworten“ enthalten. Bei diesen Variablen wurde im Falle von Variablen mit numerischer sowie nominaler Kodierung mit zwei Antwortkategorien wie folgt vorgegangen: Beim Vorliegen einer numerischen Kodierung wurde die ursprüngliche Kodierung beibehalten. Im Falle einer nominalen Kodierung mit zwei Antwortkategorien wurde die Kodierung in eine Dummy-Kodierung überführt. Unabhängig von der Kodierung wurde zusätzlich eine Dummy-Variable erstellt, welche anzeigt, ob eine Person in einer bestimmten Frage eine Kategorie außerhalb der Skala angekreuzt hat. Anschließend wurde für jede Variable einzeln geprüft, ob eine festgelegte Zuordnung der Antwort zu einem bestimmten Skalenwert inhaltlich sinnvoll ist oder ob die Wahl einer zusätzlichen Antwortkategorie als fehlender Wert behandelt werden sollte. Variablen mit einer nominalen Kodierung mit mehr als zwei Antwortkategorien wurden in der Datenanalyse als Faktoren behandelt.

Prädiktoren Nach der oben beschriebenen Datenvorverarbeitung wurden 1993 Variablen als Prädiktoren für die Benchmark Experimente verwendet.

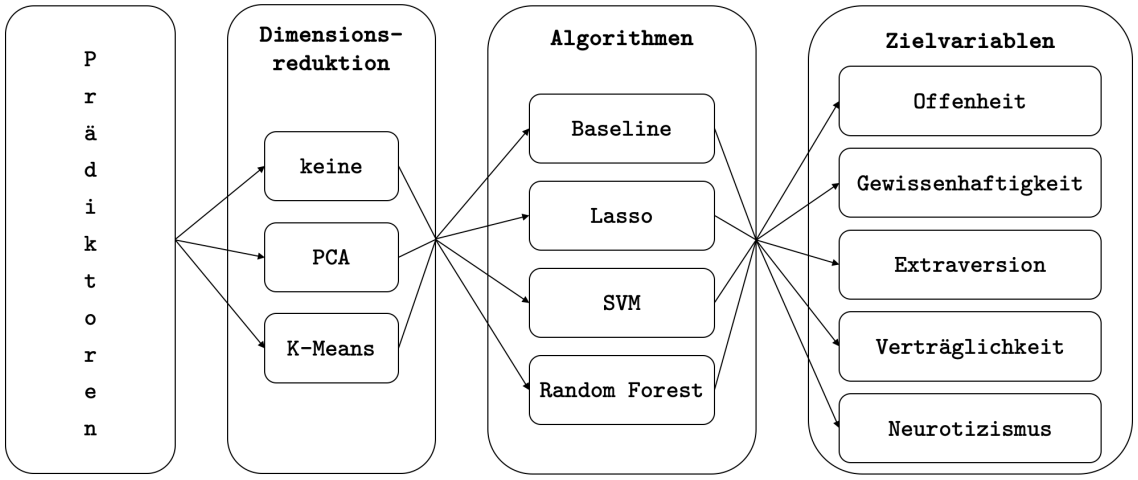
Fragen, die in mehreren Befragungswellen gestellt wurden, wurden mehrfach als Prädiktoren aufgenommen[2], da die zeitliche Zuordnung der Erhebung als informativ eingestuft wurde.

In den Prädiktoren sind Fragen zu sehr unterschiedlichen Themenbereichen enthalten. Im Folgenden werden beispielhaft einige dieser Themen aufgelistet: Einstellungen zu umweltbezogenen Themen, Parteien, Politikern, Flüchtlingen/Ausländern; Selbstbeschreibungen in Bezug auf Konsumverhalten, Vertrauen, Werte, Zeitwahr-

[2] Dabei ist an den ersten beiden Stellen der Variablenbezeichnung erkennbar, in welcher Welle die jeweilige Variable erhoben wurde.

Abbildung 3.2

Auswertungs-Design



nehmung, Naturbezug, Umweltschutz, (psychisches) Wohlbefinden, Freizeitaktivitäten, Materialismus, persönliche Motivation, Zufriedenheit in unterschiedlichen Lebensbereichen, die eigene ökonomische Situation, die eigene berufliche Tätigkeit, Einschätzungen der Notwendigkeit verschiedener Güter und Aktivitäten; Demographische Daten wie Einkommen, Beziehungs-/Familienstand, Bildungsabschluss, Geschlecht, Alter.

Kriterien Aus den Antworten auf die Fragen des BFI-10 (Rammstedt et al., 2014) in Befragungswelle *ed* wurde der Mittelwert für jede Person für die Big Five Dimensionen Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit, Neurotizismus berechnet. Der Fragebogen besteht aus zwei Fragen zu jeder Persönlichkeitsdimension, welche anhand einer fünfstufigen Likert-basierten Skala[3] bewertet wurden.

Benchmark Experimente Um zu überprüfen, ob die Kriterien aus den Prädiktoren vorhergesagt werden können, wurden Benchmark Experimente durchgeführt. Darüber hinaus dienten diese Experimente der Überprüfung, welche Modellierungen die Zusammenhänge zwischen den Prädiktoren und Kriterien am besten annähern.

Um unterschiedliche Modellierungen vergleichen zu können, wurde ein umfassendes Auswertungs-Design erstellt. Abbildung 3.2 ist eine graphische Darstellung dieses Auswertungs-Designs zu entnehmen.

Diese zeigt, dass für jedes Kriterium zwölf verschiedene Modellierungen vergli-

[3] 1= „trifft überhaupt nicht zu“, 2= „trifft eher nicht zu“, 3= „weder noch“, 4 = „eher zutreffend“, 5 = „trifft voll und ganz zu“

chen wurden. Diese unterschieden sich zunächst durch die drei verschiedenen Ansätze in der Dimensionsreduktion. Für die beiden Bedingungen, in denen eine Dimensionsreduktion durchgeführt wurde, wurde die Anzahl der Cluster bzw. der Dimensionen systematisch zwischen 1 und 25%[4] der ursprünglichen Anzahl an Variablen variiert und die mit der besten Vorhersagegüte für das finale Modell gewählt. Die unterschiedlichen Clustergrößen bzw. Anzahlen an Dimensionen wurden dabei so gewählt, dass eine Interpretation der Cluster aufgrund ihrer Anzahl noch möglich ist.

Anschließend an jede der drei Vorgehen zur Dimensionsreduktion, wurden drei verschiedene Algorithmen (SVM[5], LASSO[6] und RF[7]) zur Vorhersage des jeweiligen Kriteriums miteinander verglichen. Als Referenzalgorithmus wurde zudem ein Modell verwendet, in welchem immer der Mittelwert des Kriteriums vorhergesagt wird (Featureless). Diese Auswahl spiegelt die in anderen Studien verwendeten Algorithmen wider (Cristani et al., 2015; Farnadi et al., 2016; Guntuku et al., 2015; Lovato et al., 2014; Youyou et al., 2015). Auf Basis der Ergebnisse in Studie I (siehe Kapitel 2) wurden auf einen Entscheidungsbaum und eine lineare Modellierung ohne Regularisierung verzichtet.

Aufgrund des hohen computationalen Aufwands wurde eine 3-fach wiederholte 10-fache Kreuzvalidierung durchgeführt. Für das Hyperparametertuning in der inneren Schlefe wurde eine 5-fache Kreuzvalidierung verwendet. Im Rahmen der Kreuzvalidierung wurden für fehlende Werte in numerische Variablen der Median imputiert und in kategorialen Variablen der Modalwert. Um die Reduktion der Prädiktoren wie den Rest der Vorhersage ebenfalls kreuzvalidieren zu können und die verschiedenen Parameter-Einstellungen miteinander vergleichen zu können, wurden die genannten Methoden in eine Preproces-sing-Pipeline programmiert (Binder, 2019)[8].

Als Kriterium der Vorhersagegüte wurden verschiedene Maße betrachtet: R^2_C, die Standardabweichung des $R^2_C s$ $SD(R^2_C)$, der mittlere quadrierte Vorhersagefehler (MSE), Spearman-Rho (ρ) und die Rechenzeit (t). Eine Beschreibung dieser Maße ist Kapitel 1.3.1 zu entnehmen.

---

[4] Dabei wurden die folgenden Werte angenommen 1%, 5%, 10%, 15%, 20%, 25%

[5] Tuning Einstellungen: scale = TRUE, grid_choice = 1

[6] Tuning anhand der Grundeinstellungen des Pakets glmnet (Friedman et al., 2010) für den Befehl cv.glmnet

[7] Tuning Einstellungen: mtry = 1/3 der Prädiktorenanzahl

[8] Da der K-Means Algorithmus teilweise nicht konvergiert ist, wurden nicht konvergierende Modelle mit anderen Seeds erneut gerechnet bis eine Lösung für dieses Modell gefunden wurde.

3.3 Ergebnisse

3.3.1 Deskriptive Ergebnisse

Eine Tabelle mit deskriptiven Statistiken der Kriterien sind Tabelle B.1 (Anhang B) zu entnehmen. Aufgrund der Masse an Prädiktoren wurde von einer Tabelle mit deskriptiven Statistiken der Prädiktoren abgesehen.

Darüber hinaus wurden Spearman-Korrelationen zwischen den Prädiktoren und den Kriterien berechnet. Tabellen 3.1, 3.2, 3.3, 3.4 und 3.5 sind die Korrelationen der Prädiktoren mit Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit und Neurotizismus zu entnehmen. Im Folgenden werden die Ergebnisse der Korrelationsanalysen nach den Persönlichkeitseigenschaften geordnet beschrieben.

Offenheit

Acht Variablen stammen aus dem Bereich *Werte*. Dabei sind drei Mal die gleichen Items enthalten, welche in zwei verschiedenen Wellen erhoben wurden. Inhaltlich decken die Variablen die folgenden Bereiche ab: Wissensdurst (Offenheit für Ideen), Naturbezug, Toleranz (Offenheit für Werte- und Normensysteme), Erfahrungshunger (Offenheit für Handlungen). Generell sind die Korrelationskoeffizienten als klein anzusehen (Cohen, 1988).

Gewissenhaftigkeit

Fünf Variablen sind dem Bereich *Willenskraft* zuzuordnen, wobei zwei weitere Variablen den Begriff *Willenskraft* enthalten. Inhaltlich decken die Items die folgenden Bereiche ab: Leistungsstreben, Selbstdisziplin, Pflichtbewusstsein, Besonnenheit. Die Korrelation mit Item edbt042a („Ich bin faul“) ist hoch, während die weiteren Korrelationen eine mittlere Höhe aufweisen (Cohen, 1988).

Extraversion

Sechs der zehn Variablen enthalten Begriffe mit *Energie*, wobei bei einigen Items die Beantwortung in mehreren Wellen enthalten ist. Inhaltlich werden durch die Items die folgenden Bereiche abgedeckt: Positive Emotionen, Aktivität, Geselligkeit. Die Höhe der Korrelationen mit Extraversion ist gering (Cohen, 1988).

Verträglichkeit

Fünf der zehn Variablen befassen sich mit Einstellungen zu Flüchtlingen, zwei weitere mit der Toleranz verschiedener Menschen. Inhaltlich sind die Items den folgen-

Tabelle 3.1

Top 10 Spearman-Korrelationen zwischen den Prädiktoren und Offenheit

	Beschreibung	ID	ρ	*p*	95% CI		*N*
1	Werte: Wissen erweitern	ddze018a	.28	.000	.21	.35	2941
2	Werte: Wissen erweitern	edze018a	.28	.000	.21	.35	2934
3	Zustimmung: Kauf Öko-Produkte Teil des Lebensstils	eabk092a	.26	.000	.19	.33	2941
4	Werte: Toleranz gegenüber vielen verschiedenen Menschen	edze015a	.25	.000	.18	.32	2935
5	Werte: Sich um die Natur kümmern	edze011a	.25	.000	.18	.32	2948
6	Werte: Neue Erfahrungen machen	edze020a	.25	.000	.18	.32	2934
7	Werte: Dingen selbst auf den Grund gehen	edze027a	.25	.000	.17	.32	2947
8	Freizeitaktivität: Kenntnisse erwerben oder weiterentwickeln	dezg092a	.24	.000	.17	.31	2906
9	Werte: Neue Erfahrungen machen	ddze020a	.24	.000	.16	.31	2929
10	Werte: Sich um die Natur kümmern	ddze011a	.24	.000	.16	.31	2950

Anmerkungen. *p*-Werte und Konfidenzintervalle Bonferroni-korrigiert. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde.

den Bereichen zuzuordnen: Vertrauen, Altruismus, Gutherzigkeit, Entgegenkommen. Allgemein ist die Höhe der Korrelationen klein (Cohen, 1988).

Neurotizismus

Sechs der Variablen befassen sich damit, dass die Teilnehmer sich in den letzten Wochen ängstlich, zwei der Variablen, dass sie sich traurig gefühlt haben. Inhaltlich sind die Items den folgenden Bereichen zuzuordnen: Ängstlichkeit, Depression, Verletzlichkeit. Allgemein zeigen die Korrelationen einen mittleren Zusammenhang (Cohen, 1988).

Tabelle 3.2

Top 10 Spearman-Korrelationen zwischen den Prädiktoren und Gewissenhaftigkeit

	Beschreibung	ID	ρ	*p*	95% CI		*N*
1	Willenskraft: Ich bin faul	edbt042a	-.70	.000	-.74	-.66	2934
2	Willenskraft: Vergnügen hindert mich an Arbeit	edbt046a	-.41	.000	-.47	-.35	2945
3	Willenskraft: Eiserne Selbstdisziplin	edbt052a	.37	.000	.31	.44	2937
4	Willenskraft: Wunsch nach mehr Selbstdisziplin	edbt045a	-.36	.000	-.42	-.29	2944
5	Zeitwahrnehmung4: Erledige Vorhaben termingerecht	ddac043a	.35	.000	.28	.42	2947
6	Ökonomische Motivation: Habe keinen Antrieb Arbeit zu erledigen	debl233a	-.32	.000	-.39	-.25	2928
7	Ökonomische Motivation: Einfacher ein Projekt zu beginnen als abzuschließen	debl231a	-.32	.000	-.38	-.25	2929
8	Selbstbeschreibung: Keine Willenskraft	ebbh155a	-.30	.000	-.37	-.23	2914
9	Willenskraft: Schlechte Dinge tun	edbt044a	-.30	.000	-.37	-.23	2931
10	Selbstbeschreibung: Keine Willenskraft	debh190a	-.30	.000	-.36	-.23	2933

Anmerkungen. *p*-Werte und Konfidenzintervalle Bonferroni-korrigiert. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde.

Tabelle 3.3

Top 10 Spearman-Korrelationen zwischen den Prädiktoren und Extraversion

	Beschreibung	ID	ρ	*p*	95% CI		*N*
1	In den letzten vier Wochen glücklich gefühlt	ebaw232a	.25	.000	.18	.32	2946
2	In den letzten vier Wochen jede Menge Energie verspürt	ddaw177a	.24	.000	.17	.31	2945
3	Wichtig im Leben: Freunde und Bekannte	ebzc033a	.24	.000	.17	.31	2938
4	Freizeitaktivität: Nützliche Kontakte knüpfen	dezg090a	.24	.000	.17	.31	2904
5	Wichtigkeit: Freunde	eazb010a	.24	.000	.17	.31	2926
6	In den letzten vier Wochen jede Menge Energie verspürt	dcaw182a	.23	.000	.16	.30	2941
7	In den letzten vier Wochen energiegeladen gefühlt	ebaw235a	.23	.000	.16	.30	2940
8	In den letzten vier Wochen jede Menge Energie verspürt	ebaw243a	.23	.000	.16	.30	2924
9	Zufriedenheit: Freunde	eazb017a	.23	.000	.16	.30	2892
10	In den letzten vier Wochen jede Menge Energie verspürt	eaaw144a	.23	.000	.16	.30	2938

Anmerkungen. *p*-Werte und Konfidenzintervalle Bonferroni-korrigiert. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde.

Tabelle 3.4

Top 10 Spearman-Korrelationen zwischen den Prädiktoren und Verträglichkeit

	Beschreibung	ID	ρ	*p*	95% CI		*N*
1	Allgemeines Vertrauen	ebzc059a	.29	.000	.22	.36	2957
2	Werte: Toleranz gegenüber vielen verschiedenen Menschen	edze015a	.24	.000	.16	.31	2937
3	Flüchtlinge: Gefühle allgemein	ebbd112a	.21	.000	.14	.29	2798
4	Präferenz Flüchtlinge im Wohnumfeld	eazj114a	.21	.000	.14	.28	2951
5	Werte: Toleranz gegenüber vielen verschiedenen Menschen	ddze015a	.21	.000	.14	.28	2938
6	Willenskraft: Unangemessene Dinge sagen	edbt043a	-.21	.000	-.28	-.13	2915
7	Bedeutung Flüchtlinge für persönliche Werte	eazj109a	.20	.000	.13	.27	2937
8	Sympathie für Flüchtlinge	eazj108a	.20	.000	.13	.27	2933
9	Flüchtlinge: Gefühle allgemein	debd147a	.20	.000	.12	.27	2766
10	Flüchtlinge: Bewertung insgesamt	ebbd108a	.20	.000	.12	.27	2786

Anmerkungen. *p*-Werte und Konfidenzintervalle Bonferroni-korrigiert. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde.

Tabelle 3.5

Top 10 Spearman-Korrelationen zwischen den Prädiktoren und Neurotizismus

	Beschreibung	ID	ρ	*p*	95% CI		*N*
1	In den letzten vier Wochen ängstlich gefühlt	eaaw132a	.37	.000	.30	.43	2933
2	In den letzten vier Wochen ängstlich gefühlt	ebaw231a	.36	.000	.29	.42	2929
3	In den letzten vier Wochen ängstlich gefühlt	ddaw165a	.36	.000	.29	.42	2947
4	Willenskraft: Schwer zu konzentrieren	edbt047a	.35	.000	.28	.41	2948
5	In den letzten vier Wochen ängstlich gefühlt	deaw255a	.34	.000	.28	.40	2928
6	In den letzten vier Wochen traurig gefühlt	ebaw233a	.34	.000	.28	.41	2937
7	In den letzten vier Wochen traurig gefühlt	eaaw134a	.34	.000	.27	.41	2945
8	Zustand: Erschöpft vs. Voller Energie	edbt053a	-.34	.000	-.41	-.27	2718
9	In den letzten vier Wochen ängstlich gefühlt	dcaw170a	.34	.000	.27	.40	2929
10	In den letzten vier Wochen ängstlich gefühlt	dfaw102a	.34	.000	.27	.40	2945

Anmerkungen. *p*-Werte und Konfidenzintervalle Bonferroni-korrigiert. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde.

3.3.2 Ergebnisse der Prädiktiven Modellierung

Alle Benchmark Experimente zeigen, dass die Modellierungen ohne Dimensionsreduktion bessere als[9] bzw. gleich gute[10] Vorhersagen machen wie die Modellierungen mit Dimensionsreduktion.

Offenheit

Die Ergebnisse zur Vorhersage von Offenheit sind Tabelle 3.6 zu entnehmen. Die höchste Vorhersagegüte wurde ohne Dimensionsreduktion mit der SVM erreicht ($R^2_C = 0.25$, $SD = 0.03$) gefolgt vom LASSO ($R^2_C = 0.24$, $SD = 0.03$). Da das LASSO eine höhere Interpretierbarkeit aufweist, wird dieses Modell bevorzugt.

Tabelle 3.6

Ergebnisse des Benchmark Experiments zur Vorhersage von Offenheit

Dimensions-reduktion	Learner	R^2_C	*SD*	*t*	ρ	*MSE*
keine	Featureless	-.00	.01	3.74	NA	0.75
	SVM	.25	.03	126.68	.49	0.56
	LASSO	.24	.03	134.40	.49	0.56
	Random Forest	.17	.03	26.04	.44	0.62
PCA	Featureless	-.00	.01	1532.85	NA	0.75
	SVM	.22	.04	3773.46	.47	0.58
	LASSO	.20	.04	1551.09	.46	0.59
	Random Forest	.14	.03	1720.73	.39	0.64
K-Means	Featureless	-.00	.01	4950.29	NA	0.75
	SVM[a]	.21	.04	6991.96	.46	0.58
	LASSO[a]	.19	.03	5144.98	.44	0.60
	Random Forest[a]	.17	.03	5416.11	.43	0.62

Anmerkungen. *SD* = Standardabweichung des R^2_C, *t* = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error. [a]auf initialem Seed nicht konvergiert

Tabelle 3.7 sind die zehn Variablen mit den höchsten absoluten Regressionsgewichten zu entnehmen. Die Items können folgenden Inhalten zugeordnet werden: Offenheit für Ästhetik, Ideen sowie Gefühle. Darüber hinaus wird ein geringer Bildungsstandard als Variable beibehalten (4).

[9] Bei den Vorhersagen für Offenheit, Gewissenhaftigkeit, Extraversion und Verträglichkeit
[10] Bei der Vorhersage von Neurotizismus

Tabelle 3.7

Regressionsgewichte der Top 10 Prädiktoren zur Vorhersage von Offenheit mit einem LASSO

	Beschreibung	ID	*b*
	(Intercept)		1.42
1	Beteiligung in Organisationen: Verein für Kunst, Musik, Kulturelles	ebzc023a	0.09
2	Naturbezug: Natur schöner als Kunstwerk	ddbk071a	-0.08
3	Werte: Wissen erweitern	ddze018a	0.06
4	Abschluss einer Verwaltungsfachhochschule[a]	dfzh047a	-0.05
5	Polareis Wissen über Abschmelzen	edzy054a	0.05
6	Naturbezug: Sammle Souvenirs aus Natur	ddbk074a	0.05
7	Lebensstandard Notwendigkeit: Kino, Theater oder Konzert[b]	ddbg093a	-0.05
8	Zeitwahrnehmung1: Vertraute Bilder, Geräusche wecken Erinnerungen	ddac029a	0.04
9	Werte: Dingen selbst auf den Grund gehen	edze027a	0.04
10	Freizeitaktivität: Kenntnisse erwerben oder weiterentwickeln	dezg092a	0.04

Anmerkungen. [a]Dummy Variable, welche angibt, dass die Personen einen Abschluss einer Verwaltungsfachhochschule haben. [b]Antwortformat: 1 (absolut notwendig) bis 3 (verzichtbar). Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde. Insgesamt wurden 80 Prädiktoren für das Modell beibehalten.

Gewissenhaftigkeit

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Gewissenhaftigkeit sind Tabelle 3.8 zu entnehmen. Die höchste Vorhersagegüte wird durch eine Modellierung ohne Dimensionsreduktion mit einem LASSO ($R^2_C = 0.54$, $SD = 0.04$) erreicht.

Tabelle 3.9 sind die zehn höchsten Regressionskoeffizienten der Variablen zu entnehmen, die im LASSO beibehalten wurden. Die Items können folgenden Inhalten zugeordnet werden: Leistungsstreben, Selbstdisziplin, Pflichtbewusstsein und Besonnenheit. Darüber hinaus wird die Variable Geschlecht beibehalten sowie eine Variable, die Altruismus (Verträglichkeit) beschreibt.

Tabelle 3.8

Ergebnisse des Benchmark Experiments zur Vorhersage von Gewissenhaftigkeit

Dimensions-reduktion	Learner	R^2_C	*SD*	*t*	ρ	*MSE*
keine	Featureless	-.00	.00	3.43	NA	0.51
	SVM	.49	.04	120.73	.70	0.26
	LASSO	.54	.04	102.34	.75	0.23
	Random Forest	.38	.03	25.32	.69	0.31
PCA	Featureless	-.00	.00	1530.14	NA	0.51
	SVM	.45	.05	3478.27	.67	0.28
	LASSO	.44	.04	1554.77	.66	0.28
	Random Forest	.26	.03	1718.61	.58	0.38
K-Means	Featureless[a]	-.00	.00	5106.80	NA	0.51
	SVM[a]	.43	.06	6981.05	.65	0.29
	LASSO	.42	.05	5128.61	.64	0.29
	Random Forest[a]	.36	.05	5402.46	.63	0.33

Anmerkungen. SD = Standardabweichung des R^2_C, t = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error. [a]auf initialem Seed nicht konvergiert

Extraversion

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Extraversion können Tabelle 3.10 entnommen werden. Die höchste Vorhersagegüte konnte ohne Dimensionsreduktion mit der SVM erreicht werden ($R^2_C = 0.23$, $SD = 0.05$), während das LASSO eine vergleichbare Vorhersagegüte erreichte ($R^2_C = 0.21$, $SD = 0.04$). Aufgrund der besseren Interpretierbarkeit wird daher das LASSO bevorzugt.

Tabelle 3.11 sind die höchsten zehn Regressionskoeffizienten der Variablen zu entnehmen, die im LASSO beibehalten wurden. Die Items können folgenden Inhalten zugeordnet werden: Geselligkeit, Durchsetzungsfähigkeit sowie positive Emotionen. Darüber hinaus wird eine Variable, die Leistungsstreben (Gewissenhaftigkeit) zuzuordnen ist und die Variable Geschlecht beibehalten. Auffällig ist zudem, dass zwei Variablen enthalten sind, welche die Dummy-Kodierung der gewählten Kategorie „weiß nicht“ repräsentieren.

Verträglichkeit

Die Ergebnisse des Benchmark Experiments zur Vorhersage von Verträglichkeit sind Tabelle 3.12 zu entnehmen. Ohne Dimensionsreduktion wird mit der SVM sowie

Tabelle 3.9

Regressionsgewichte der Top 10 Prädiktoren zur Vorhersage von Gewissenhaftigkeit mit einem LASSO

	Beschreibung	ID	*b*
	(Intercept)		3.97
1	Willenskraft: Ich bin faul	edbt042a	-0.36
2	Willenskraft: Eiserne Selbstdisziplin	edbt052a	0.07
3	Zeitwahrnehmung4: Erledige Vorhaben termingerecht	ddac043a	0.06
4	Ökonomische Motivation: Einfacher ein Projekt zu beginnen als abzuschließen	debl231a	-0.05
5	Willenskraft: Vergnügen hindert mich an Arbeit	edbt046a	-0.04
6	Geschlecht[a]	dfzh037a	-0.02
7	Werte: Menschen helfen, die einem am Herzen liegen	edze019a	0.02
8	Zeitwahrnehmung1: Wenn ich etwas erreichen will, setzte ich mir Ziele	ddac033a	0.02
9	Zeitwahrnehmung2: Wichtiger zu genießen, was man gerade tut	ddac036a	-0.02
10	Werte: Alle Gesetze befolgen	edze022a	0.02

Anmerkungen. [a]Männlich = 1, weiblich = 0. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde. Insgesamt wurden 31 Prädiktoren im Modell beibehalten

einem LASSO die gleiche Vorhersagegüte erreicht ($R^2_C = 0.20$, $SD = 0.03$). Aufgrund der besseren Interpretierbarkeit wird das LASSO bevorzugt.

Tabelle 3.13 sind die zehn höchsten Regressionskoeffizienten der Variablen zu entnehmen, die im LASSO beibehalten werden. Die Items können folgenden Inhalten zugeordnet werden: Entgegenkommen, Vertrauen, Altruismus und Gutherzigkeit. Auffällig ist zudem, dass drei Variablen enthalten sind, welche die Dummy-Kodierung der gewählten Kategorie „weiß nicht“ repräsentieren.

Neurotizismus

Tabelle 3.14 sind die Ergebnisse des Benchmark Experiments zur Vorhersage von Neurotizismus zu entnehmen. Die höchste Vorhersagegüte wird ohne Dimensionsreduktion mit einer SVM erreicht ($R^2_C = 0.32$, $SD = 0.05$). Eine vergleichbar hohe Vohersagegüte wird durch die Modellierung ohne Dimensionsreduktion mit einem LASSO ($R^2_C = 0.31$, $SD = 0.04$) sowie mit einer Dimensionsreduktion durch eine PCA mit einer SVM ($R^2_C = 0.31$, $SD = 0.05$) erreicht. Aufgrund der einfacheren Interpretierbarkeit wird das Modell ohne Dimensionsreduktion mit dem LASSO be-

Tabelle 3.10

Ergebnisse des Benchmark Experiments zur Vorhersage von Extraversion

Dimensions-reduktion	Learner	R^2_C	*SD*	*t*	ρ	*MSE*
keine	Featureless	-.00	.01	3.42	NA	0.74
	SVM	.23	.05	128.29	.46	0.57
	LASSO	.21	.04	135.89	.45	0.59
	Random Forest	.15	.03	26.33	.40	0.63
PCA	Featureless	-.00	.01	1536.21	NA	0.74
	SVM	.20	.04	3795.78	.43	0.59
	LASSO	.17	.03	1562.03	.41	0.61
	Random Forest	.12	.02	1737.83	.38	0.65
K-Means	Featureless[a]	-.00	.01	5041.32	NA	0.74
	SVM	.18	.04	6924.00	.41	0.61
	LASSO[a]	.16	.04	5302.95	.40	0.62
	Random Forest[a]	.14	.03	5403.09	.38	0.64

Anmerkungen. SD = Standardabweichung des R^2_C, *t*= Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error. [a]auf initialem Seed nicht konvergiert

vorzugt.

Tabelle 3.15 sind die zehn höchsten Regressionsgewichte der Variablen zu entnehmen, die im LASSO beibehalten wurden. Die Items können folgenden Inhalten zugeordnet werden: Verletzlichkeit, Impulsivität, Soziale Befangenheit, Depression und Ängstlichkeit. Darüber hinaus wird die Variable Geschlecht beibehalten.

3.4 Diskussion

In der vorliegenden Studie wurde explorativ untersucht, inwiefern vorhandene Daten aus Panel-Befragungen im Rahmen der Validierung von Fragebögen genutzt werden können. Hierzu wurde ein Datensatz aus dem GESIS-Panel verwendet (GESIS, 2018). Dabei wurde die Validität des BFI-10 (Rammstedt et al., 2014) untersucht, welcher bereits im verwendeten Datensatz enthalten war. Nach Aggregation der Itemantworten zu Skalen für die jeweilige Persönlichkeitsdimension wurden einerseits Korrelationsanalysen gerechnet, andererseits ML Algorithmen verwendet, um die Big Five Persönlichkeitseigenschaften vorherzusagen. Für beide Methoden wurden die Variablen mit den höchsten Zusammenhängen mit dem jeweiligen Kriterium betrachtet. Die Ergebnisse zeigen, dass sowohl die Korrelationsanalysen als

Tabelle 3.11

Regressionsgewichte der Top 10 Prädiktoren zur Vorhersage von Extraversion mit einem LASSO

	Beschreibung	ID	b
	(Intercept)		1.64
1	Wichtig im Leben: Freunde und Bekannte	ebzc033a	0.09
2	Lebensstandard Verfügbarkeit: Fernseher: Habe bzw. tue ich nicht[a]	ddbg136a	-0.09
3	Alkoholkonsum: Weiß nicht[b]	ebbm220a	-0.08
4	Willenskraft: Ich bin faul	edbt042a	-0.08
5	Geschlecht[c]	dfzh037a	-0.08
6	Freizeitaktivität: Nützliche Kontakte knüpfen	dezg090a	0.07
7	Kontakte mit Freunden	ebzc031a	0.06
8	Zeitwahrnehmung4: Erfreuliche Erfahrungen kommen leicht in den Sinn	ddac044a	0.06
9	ALLBUS: Aktive Rolle bei politischen Fragen[d]	ecbo085a	-0.06
10	Flüchtlinge auflagenlose Arbeitserlaubnis: Weiß ich nicht[b]	dczy158a	-0.05

Anmerkungen. [a]Dummy Variable, welche angibt, dass die Personen keinen Fernseher haben (andere Antwortmöglichkeiten waren, dass sie einen haben oder sich keinen leisten können). [b]Dummy Variable erstellt aus der Antwortkategorie „weiß nicht". [c]Männlich = 1, weiblich = 0. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde. [d] Item negativ gepolt.
Insgesamt wurden 125 Prädiktoren im Modell beibehalten.

auch die Vorhersagen der Persönlichkeitseigenschaften informativ für die Validität des Fragebogens sind.

In Bezug auf die Modellierung durch ML Algorithmen zeigen die Ergebnisse, dass lineare Modelle zur Vorhersage von Persönlichkeitseigenschaften zu bevorzugen sind. Darüber hinaus wird deutlich, dass eine Dimensionsreduktion in einem Vorverarbeitungsschritt keine Verbesserung der Vorhersage bringt. Dies ist in Einklang mit den Ergebnissen von Farnadi et al. (2016), die in ihrer Studie ebenfalls keine Steigerung der Vorhersagegüte durch Variablenselektion zeigen konnten.

3.4.1 Diskussion der Studienergebnisse

Um zu evaluieren, inwiefern ML Algorithmen im Rahmen der Fragebogenentwicklung zur Validierung eines Fragebogens verwendet werden können erfolgt zunächst eine inhaltliche Interpretation der Ergebnisse. Dabei wird überprüft, inwiefern es

Tabelle 3.12

Ergebnisse des Benchmark Experiments zur Vorhersage von Verträglichkeit

Dimensions-reduktion	Learner	R^2_C	*SD*	*t*	ρ	*MSE*
keine	Featureless	-.00	.01	3.43	NA	0.49
	SVM	.20	.03	126.68	.44	0.39
	LASSO	.20	.03	139.38	.45	0.39
	Random Forest	.14	.03	26.31	.40	0.43
PCA	Featureless	-.00	.01	1529.58	NA	0.49
	SVM	.18	.04	3620.04	.42	0.40
	LASSO	.16	.03	1557.02	.40	0.41
	Random Forest	.11	.02	1739.13	.35	0.44
K-Means	Featureless[a]	-.00	.01	5047.10	NA	0.49
	SVM[a]	.18	.03	6954.26	.42	0.40
	LASSO	.16	.03	5184.53	.41	0.41
	Random Forest[a]	.14	.03	5316.08	.39	0.42

Anmerkungen. *SD* = Standardabweichung des R^2_C, *t* = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error. [a]auf initialem Seed nicht konvergiert

für die Variablen, die im Rahmen der verschiedenen Analysen den höchsten Zusammenhang mit den Kriterien zeigten, theoretische Erklärungen gibt.

Daran anschließend werden die Ergebnisse in Bezug auf die übergeordnete Fragestellung diskutiert.

Offenheit Die Items, welche die höchsten Korrelationen bzw. die höchsten Regressionsgewichte mit Offenheit aufweisen, können durch entsprechende Forschungsergebnisse gestützt werden.

Der Inhalt vieler Items ist im Einklang mit der theoretischen Konzeption von Offenheit. So wird Offenheit in einigen Publikationen als Intellekt bezeichnet (z. B. McCrae & John, 1992; Saucier & Goldberg, 1996; Woo et al., 2014), was sich in den Variablen durch das Streben nach Wissen widerspiegelt. Auch kulturelles Interesse wird als Teil von Offenheit beschrieben (z. B. McCrae & John, 1992), was in den Variablen durch Beteiligung in kulturellen Vereinen sowie den Besuch kultureller Veranstaltungen repräsentiert ist. Darüber hinaus ist Neugierde ein Teil von Offenheit (Saucier & Goldberg, 1996; Woo et al., 2014), was in den Variablen durch das Interesse, Dinge zu verstehen und neue Erfahrungen zu machen enthalten ist. Letztlich wird auch Toleranz von Woo et al. (2014) als Facette von Offenheit beschrieben.

Tabelle 3.13

Regressionsgewichte der Top 10 Prädiktoren zur Vorhersage von Verträglichkeit mit einem LASSO

	Beschreibung	ID	*b*
	(Intercept)		2.84
1	Willenskraft: Unangemessene Dinge sagen	edbt043a	-0.08
2	Allgemeines Vertrauen	ebzc059a	0.05
3	Werte: Toleranz gegenüber vielen verschiedenen Menschen	edze015a	0.05
4	ALLBUS: Konflikt zwischen Ausländern und Deutschen: Weiß nicht[a]	ecbo095a	0.03
5	ALLBUS: Konflikt zwischen politisch links und politisch rechts: Weiß nicht[a]	dfbo069a	0.03
6	Werte: Anderen sagen was sie tun sollen	ddze021a	-0.03
7	NEP-Skala: Menschlicher Einfallsreichtum[b]	eczd005a	-0.03
8	Werte: Sich immer eine eigene Meinung zu bilden	edze013a	-0.03
9	ALLBUS: Rangunterschiede zwischen Menschen akzeptabel: Weiß nicht[a]	dfbo066a	0.03
10	Wichtig im Leben: Religion	ebzc037a	0.02

Anmerkungen. [a]Dummy Variable erstellt aus der Antwortkategorie „weiß nicht". [b]Item negativ gepolt. Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde.
Insgesamt wurden 70 Prädiktoren im Modell beibehalten.

Dass Bilder der Geräusche positive Erinnerungen wecken (ddac029a), könnte durch die Facette *Offenheit für Gefühle* (Ostendorf & Angleitner, 2004; Woo et al., 2014) erklärt werden, da diese beschreibt, dass offene Personen intensive und vielfältige Emotionen erleben.

Des Weiteren können Befunde aus weiteren Studien die Relevanz der Variablen erklären. So konnte in verschiedenen Studien gezeigt werden, dass umweltfreundliches Verhalten mit Offenheit zusammenhängt (Brick & Lewis, 2016; Hilbig, Zettler, Moshagen & Heydasch, 2013). Zusammen mit dem Streben nach Wissen könnte dies auch erklären, dass offene Personen über das Abschmelzen des Polareises informiert sind.

Außerdem kann Natur bei offenen Personen Faszination auslösen (Silvia, Fayn, Nusbaum & Beaty, 2015), was erklären könnte, dass sie diese schöner finden als Kunst und entsprechende Souvenirs sammeln.

Tabelle 3.14

Ergebnisse des Benchmark Experiments zur Vorhersage von Neurotizismus

Dimensions-reduktion	Learner	R^2_C	*SD*	*t*	ρ	*MSE*
keine	Featureless	-.00	.00	3.47	NA	0.65
	SVM	.32	.05	126.87	.55	0.44
	LASSO	.31	.04	132.23	.54	0.45
	Random Forest	.25	.03	25.93	.51	0.49
PCA	Featureless	-.00	.00	1531.54	NA	0.65
	SVM	.31	.05	3697.95	.53	0.45
	LASSO	.30	.04	1561.18	.53	0.46
	Random Forest	.22	.03	1742.18	.48	0.51
K-Means	Featureless[a]	-.00	.00	5046.89	NA	0.65
	SVM	.30	.05	6899.59	.52	0.46
	LASSO[a]	.28	.04	5171.93	.52	0.47
	Random Forest[a]	.25	.04	5388.41	.50	0.49

Anmerkungen. *SD* = Standardabweichung des R^2_C, *t* = Rechenzeit pro Modell in Sekunden, ρ = Spearman-Rho, *MSE* = Mean Squared Error. [a]auf initialem Seed nicht konvergiert

Gewissenhaftigkeit Auch die Variablen, welche die höchsten Korrelationen bzw. Regressionsgewichte mit Gewissenhaftigkeit aufweisen, können zu einem großen Teil durch die theoretische Konzeption des Konstrukts erklärt werden.

Gewissenhafte Personen werden als selbstdiszipliniert (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996) beschrieben, die pünktlich sind (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), sodass es nahe liegt, dass sie ihre Aufgaben termingerecht erledigen. Sie werden als weitsichtig und planvoll (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996) beschrieben, sodass sie sich Ziele setzen, die sie erreichen wollen. Darüber hinaus sind sie rechtschaffen (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), weshalb sie Gesetze befolgen.

Wenig gewissenhafte Personen werden als faul (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996) beschrieben, was mit wenig Willenskraft und somit Selbstdisziplin sowie wenig Antrieb, Arbeit zu erledigen, einhergehen könnte. Auch die Fähigkeit, Projekte einfacher zu beginnen als abzuschließen kann mit geringer Beharrlichkeit (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996) erklärt werden. Darüber hinaus werden sie als hedonistisch beschrieben (Ostendorf & Angleitner, 2004) was erklären könnte, dass Vergnügen gegenüber Arbeit bevorzugt wird und die Personen angeben, manchmal Dinge zu tun, die ihnen Spaß machen,

Tabelle 3.15

Regressionsgewichte der Top 10 Prädiktoren zur Vorhersage von Neurotizismus mit einem LASSO

	Beschreibung	ID	*b*
	(Intercept)		2.97
1	Willenskraft: Schwer zu konzentrieren	edbt047a	0.10
2	Geschlecht[a]	dfzh037a	-0.09
3	Aktuelle Zugehörigkeit Kirche/ Religionsgemeinschaft: Das möchte ich nicht beantworten[b]	edzt028a	-0.07
4	Lebensstandard Verfügbarkeit: Waschmaschine: Habe bzw. tue ich nicht[c]	ddbg135a	0.06
5	Willenskraft: Unerschöpflich	edbt037a	-0.06
6	Zeitwahrnehmung3: Ein Leben ohne Risiko zu langweilig	ddac041a	-0.04
7	Willenskraft: Alternativen nicht durchdacht	edbt050a	0.04
8	Werte: Sich immer eine eigene Meinung zu bilden	edze013a	-0.04
9	Willenskraft: Volle Kraft für weitere Aktivitäten	edbt039a	-0.03
10	Willenskraft: Effektiv auf Ziele hinarbeiten	edbt048a	-0.03

Anmerkungen. [a]Männlich = 1, weiblich = 0 [b]Dummy-Variable, erstellt aus der Antwortkategorie *Das möchte ich nicht beantworten.* [c]Dummy Variable, welche angibt, dass die Personen keine Waschmaschine haben (Andere Antwortmöglichkeiten waren, dass sie eine haben oder sich keine leisten können). Die Beschreibung der Variablen wurde aus dem Handbuch des GESIS Panels entnommen. Die ID ermöglicht die Zuordnung der Variablen, wobei die ersten beiden Buchstaben die Welle angeben, in der das Item erhoben wurde. Insgesamt wurden 92 Prädiktoren im Modell beibehalten.

obwohl sie für sie schlecht sind.

Für Gewissenhaftigkeit konnte gezeigt werden, dass Frauen in entwickelten Ländern höhere Werte erreichen als Männer (Schmitt, Realo, Voracek & Allik, 2008). Für den Zusammenhang zwischen Hilfeverhalten gegenüber nahestehenden Personen und Gewissenhaftigkeit konnten in der Literatur keine Befunde gefunden werden.

Extraversion Unter den Variablen, die die höchsten Korrelationen bzw. Regressionsgewichte mit Extraversion aufweisen, entsprechen viele der theoretischen Konzeption des Konstrukts.

Extravertierte Personen, werden als glücklich beschrieben (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996). Sie sind energetisch (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996) und gesellig (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), weshalb ihnen der Kontakt mit anderen Personen wichtig ist. Da Fernsehen eine wenig gesellige Freizeitaktivität ist, könnte auch erklären, warum

sie keinen Fernseher besitzen. Sie erleben häufig positive Emotionen (Ostendorf & Angleitner, 2004), was erklären könnte, warum sie sich leicht an erfreuliche Erfahrungen erinnern. Darüber hinaus sind extravertierte Personen durchsetzungsfähig (Ostendorf & Angleitner, 2004), was erklären könnte, warum sie in politischen Fragen eine aktive Rolle übernehmen und eher eine Meinung dazu gebildet haben, ob Flüchtlinge eine auflagenlose Aufenthaltserlaubnis haben sollten.

Für Extraversion wurde zudem gezeigt, dass Frauen in entwickelten Ländern höhere Werte erreichen als Männer (Schmitt et al., 2008).

In Bezug auf den Alkoholkonsum wurde gezeigt, dass Extraversion mit Trinkverhalten zusammenhängt (Martsh & Miller, 1997). Daher scheint es nachvollziehbar, dass extravertierte Personen über ihren Alkoholkonsum Bescheid wissen.

Dafür, dass extravertierte Personen faul sind, konnten keine Belege in der Literatur gefunden werden.

Verträglichkeit Einige Variablen, die eine hohe Korrelation bzw. hohe Regressionsgewichte aufweisen, können durch die theoretische Konzeption von Verträglichkeit erklärt werden.

Personen mit hohen Verträglichkeitswerten haben ausgeprägtes Vertrauen (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996) und sind optimistisch (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), was erklären könnte, dass sie davon überzeugt sind, dass menschlicher Einfallsreichtum dazu führen wird, dass wir unsere Welt nicht zerstören[11]. Vertrauen könnte auch erklären, warum sie sich nicht immer eine eigene Meinung bilden. Sie sind mitfühlend (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), was dazu führen könnte, dass sie sich gegenüber anderen solidarisch zeigen. Dadurch, dass sie rücksichtsvoll sind (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), werden sie nur selten unangemessene Dinge sagen. Zudem sind sie nachgiebig und wenig dominierend (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), was erklären könnte, dass sie anderen ungern sagen, was sie tun sollen. Des Weiteren werden sie als religiös beschrieben (Saucier & Goldberg, 1996). Sie zeigen Toleranz (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), was auch erklären könnte, dass sie sich nicht sicher sind, inwiefern es einen Konflikt zwischen Ausländern und Deutschen bzw. politischen Lagern gibt.

[11] Item eczd005a *NEP-Skala: Menschlicher Einfallsreichtum* hat folgenden Wortlaut: Der menschliche Einfallsreichtum wird dafür sorgen, dass wir die Erde NICHT unbewohnbar machen.

Dass viele Variablen enthalten sind, welche eine positive Einstellung gegenüber Flüchtlingen und Ausländern beschreiben, kann einerseits über Toleranz erklärt werden. Darüber hinaus konnte gezeigt werden, dass Verträglichkeit damit zusammenhängt, dass Personen eine länderübergreifende soziale Orientierung aufweisen[12] (Butrus & Witenberg, 2013; Reese, Proch & Cohrs, 2014).

Neurotizismus Auch für das Kriterium Neurotizismus können einige Variablen, die eine hohe Korrelation bzw. hohe Regressionsgewichte zeigen, durch die theoretische Konzeption des Konstrukts erklärt werden.

Neurotische Personen werden als ängstlich beschrieben (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996), was erklären könnte, warum sie kein risikoreiches Leben bevorzugen. Sie neigen dazu, niedergeschlagen und traurig zu sein (Ostendorf & Angleitner, 2004; Saucier & Goldberg, 1996). Darüber hinaus werden sie als gestresst und unruhig (Ostendorf & Angleitner, 2004) beschrieben, was erklären könnte, dass sie sich als erschöpft und kraftlos beschreiben und leicht aus dem Gleichgewicht kommen. Die eingeschränkte Willenskraft kann dadurch erklärt werden, dass neurotische Personen an sich selbst zweifeln (Ostendorf & Angleitner, 2004). Darüber hinaus werden Personen, die hohe Neurotizismus Werte haben, als hilflos beschrieben (Ostendorf & Angleitner, 2004), was dazu führen kann, dass sie Alternativen nicht durchdenken, sich nicht immer eine eigene Meinung bilden und es ihnen schwerfällt, effektiv auf Ziele hinzuarbeiten.

Auch für Neurotizismus wurde gezeigt, dass Frauen in entwickelten Ländern höhere Werte zeigen als Männer (Schmitt et al., 2008).

Warum Personen, die keine Auskunft über ihre Zugehörigkeit zur Religionsgemeinschaft geben wollen, niedrigere Werte in Neurotizismus erreichen, ist nicht durch entsprechende Studien belegbar.

Ebenso ist anhand der Forschungsliteratur der Zusammenhang zum Besitz einer Waschmaschine nicht erklärbar.

Allgemeine Ergebnisse Da viele Variablen, welche die höchsten Korrelationen bzw. Regressionsgewichte zum jeweiligen Kriterium zeigten, durch die Konzeptionierung der jeweiligen Persönlichkeitseigenschaft erklärt werden können, kann davon ausgegangen werden, dass die verwendeten Verfahren Hinweise zur Inhaltsvalidierung geben. Die Items der Top 10 Prädiktoren, für die Zusammenhänge zu den Big Five in inhaltlichen Studien gezeigt werden konnten, können als Indikator dafür gese-

[12] Global Social Identification

hen werden, an welchen Stellen Zusammenhänge genauer untersucht werden sollten bzw. überprüft werden sollte, inwiefern diese Aspekte in der Konzeptionierung aufgenommen werden sollten (Shmueli, 2010). Darüber hinaus sollte besonderes Augenmerk darauf gelegt werden, für welche Variablen keine Erklärung gefunden werden konnte. Hierbei sollte allerdings berücksichtigt werden, dass in prädiktiven Modellen auch Variablen wichtig sein können, welche keinen kausalen Zusammenhang zum Kriterium aufweisen (Shmueli, 2010).

Des Weiteren sind unter den Variablen mit den stärksten Zusammenhängen mit dem jeweiligen Kriterium teilweise Dummy-Variablen enthalten, welche angeben, dass eine Person die Kategorie „weiß nicht“ angekreuzt hat. Die Verwendung solcher Variablen kann für prädiktive Modelle von Vorteil sein, während die Interpretation dessen unklar bleibt (Shmueli, 2010).

In der vorliegenden Studie ist Geschlecht unter den Variablen mit den höchsten Gewichten in den Vorhersagen von Neurotizismus, Gewissenhaftigkeit und Extraversion enthalten. Dies stimmt mit der Literatur überein, in welcher Schmitt et al. (2008) zeigen konnten, dass Frauen vor allem in entwickelten Ländern höhere Werte als Männer in den Persönlichkeitseigenschaften Neurotizismus, Extraversion, Verträglichkeit und Gewissenhaftigkeit zeigen. In der vorliegenden Studie wird im Vorhersagemodell für Verträglichkeit Geschlecht allerdings nicht im Modell beibehalten. Dementsprechend sollte überprüft werden, inwiefern dies eine relevante Abweichung gegenüber anderen Instrumenten zur Persönlichkeitsmessung darstellt.

Bleidorn und Hopwood (2019) weisen darauf hin, dass es bei der Verwendung von ML sein kann, dass Persönlichkeit nicht als stabiler Trait vorhergesagt wird, sondern eher State-basierte Konstrukte, die ähnlich zu Persönlichkeitseigenschaften sind. Dass die besten Prädiktoren für das jeweilige Kriterium unterschiedlichen Wellen zuzuordnen sind, deutet darauf hin, dass der Kurzfragebogen einen stabilen Trait misst. Würde es einer State-Messung entsprechen, so sollten vor allem Variablen aus der gleichen Welle zur Vorhersage des Kriteriums relevant sein, da der aktuelle State die Beantwortung aller Fragen einer Befragungswelle beeinflussen sollte.

3.4.2 Stärken, Limitationen und Ausblick

Verwendung von Panel-Daten Durch die Verwendung von Panel-Daten, welche eine repräsentative Stichprobe für die deutsche Bevölkerung enthalten, knüpft die vorliegende Studie an die schon lange geführte Diskussion an, dass in der psychologischen Forschung weniger Studenten-Stichproben verwendet werden sollten, um die

Generalisierbarkeit der Ergebnisse zu erhöhen (Henrich et al., 2010a, 2010b; Peterson, 2001; Rad et al., 2018; Wintre et al., 2007). Dadurch, dass die Aufteilung des Datensatzes im Rahmen des Resamplings zufällig durchgeführt wurde, kann allerdings nicht sichergestellt werden, dass die verwendeten Teildatensätze ebenfalls repräsentative Stichproben enthielten. In zukünftigen Studien sollte daher überprüft werden, inwiefern die Repräsentativität trotz der Verwendung von Resampling-Strategien erhalten bleiben kann.

Darüber hinaus bietet die Verwendung von Panel-Daten die Möglichkeit, den Fragebogen mit einer sonst nicht zugänglichen Menge an anderen Variablen zu vergleichen. Somit können umfangreiche Hinweise für die Validierung gefunden werden.

Dimensionsreduktion Die Ergebnisse der Benchmark Experimente deuten darauf hin, dass eine Dimensionsreduktion zur Vorhersage von Persönlichkeitseigenschaften keine Leistungssteigerung der ML Algorithmen bringt. Somit scheint es für zukünftige Studien nicht sinnvoll, computationale Ressourcen dafür zu verwenden. Allerdings kann diese Erkenntnis nicht uneingeschränkt auf alle Dimensionsreduktions-Verfahren verallgemeinert werden.

In der vorliegenden Studie wurden zwei Verfahren verwendet, die als nicht optimal angesehen werden können. Beide Verfahren haben den Nachteil, dass die Anzahl der zu extrahierenden Cluster bzw. Dimensionen vorgegeben werden muss (Chavent et al., 2012; Chavent et al., 2017; Jain et al., 1999; Worthington & Whittaker, 2006). Somit kann es sein, dass die verwendeten Zahlen an Clustern bzw. Dimensionen keine optimale Lösung ermöglichen (Jain et al., 1999).

Darüber hinaus werden zur besseren Interpretierbarkeit der Dimensionen bei der PCA typischerweise Rotationen durchgeführt, sodass orthogonale Dimensionen entstehen (Abdi & Williams, 2010). Dies wurde in der vorliegenden Studie vernachlässigt, da der entsprechende Algorithmus große Konvergenzprobleme zeigte.

Des Weiteren kann K-Means Clustering vor allem kompakte und gut separierbare Cluster identifizieren (Jain et al., 1999). Allerdings ist es denkbar, dass dies nicht dem entspricht, was in den verwendeten Daten gegeben ist, da einige enthaltene Variablen zueinander ähnlich, aber nicht gleich sind. Auch lief die in der vorliegenden Studie verwendete algorithmische Umsetzung des K-Means Clusterings nicht stabil und zeigte für verschiedene Konstellationen im Datensatz Konvergenzprobleme.

Als Alternative zu den in der vorliegenden Studie verwendeten Verfahren, bei denen ein Ähnlichkeitsmaß zu den entstandenen Dimensionen angegeben wird (Abdi & Williams, 2010; Chavent et al., 2012), könnte aus jedem Cluster eine Variable stellvertretend für dieses ausgewählt werden (Vigneau & Qannari, 2003). Die Ver-

wendung einer solchen stellvertretenden Variable scheint gerade bei großen Datensätzen mit vielen Variablen, welche zueinander ähnliche Konstrukte messen, interessant. Dadurch, dass in den verwendeten Panel-Daten viele Fragebögen integriert sind, könnte es sein, dass Items, welche das gleiche Konstrukt messen sollen, ein Cluster bilden. Daher könnte die Auswahl einer stellvertretenden Variable für jedes gemessene Konstrukt für Validierungsfragestellungen relevant sein.

ML und Fragebogen-Validierung Die Frage, ob ML Algorithmen zur Validierung von Fragebögen verwendet werden können, kann auf Basis der vorliegenden Ergebnisse positiv beantwortet werden. Dies ist im Einklang mit der Argumentation von Bleidorn und Hopwood (2019).

In der vorliegenden Studie zeigte das LASSO im Vergleich zu anderen Modellierungen eine gute Vorhersagegüte. Gerade lineare Modelle scheinen für Validierungsstudien sinnvoll verwendbar, da sie gut interpretierbare Modelle als Ergebnisse liefern. Vor allem durch solche Modelle kann der Forderung von Bleidorn et al. (2017) nachgekommen werden, neben der Vorhersage von Persönlichkeitseigenschaften auch die dahinter liegenden Algorithmen zu verstehen.

Gerade im Zusammenhang mit der Verwendung eines großen Datensatzes scheint es sinnvoll, neben den Korrelationen auch die Ergebnisse des LASSO zu betrachten. Im LASSO wird von Variablen, welche hohe Korrelationen untereinander aufweisen, oftmals lediglich eine beibehalten (Zou & Hastie, 2005). Das hat den Vorteil, dass unter den Variablen, welche im Modell beibehalten werden, nur wenige Variablen enthalten sind, welche die gleiche Frage in unterschiedlichen Befragungswellen repräsentieren. Somit scheint es möglich, über die im Modell beibehaltenen Variablen weniger redundante Informationen über zusammenhängende Variablen zu erhalten als Korrelationsanalysen.

Darüber hinaus wurde die Anzahl der relevanten Prädiktoren in der vorliegenden Studie von 1993 auf 31 bis 125 für die einzelnen Vorhersagen reduziert. Während es einen enormen Aufwand bedeuten würde, sich die Korrelationen zwischen 1993 Variablen und dem jeweiligen Kriterium anzuschauen und inhaltlich zu prüfen, scheint eine einzelne Betrachtung der reduzierten Prädiktoren durchaus möglich. Da der verwendete Panel-Datensatz auch Variablen enthält, die im Rahmen einer speziell zur Validierung eines Big Five Fragebogens durchgeführten Validierungsstudie nicht enthalten wären, sollten darüber hinaus die gewonnenen Informationen über die Konstrukte umfangreicher sein (Bleidorn & Hopwood, 2019).

Die hier verwendeten Methoden scheinen allerdings nur einen Teil der Inhalts-

validierung zu umfassen. Um eine umfassende Inhaltsvalidierung des Fragebogens zu erhalten, sollte neben der hier dargestellten positiven Selektion auch betrachtet werden, welche Prädiktoren, die einen theoretisch fundierten Zusammenhang zum Kriterium aufweisen sollten, diesen in den vorliegenden Daten nicht zeigen. Dies stellt besonders bei der großen Anzahl an Prädiktoren einen enormen Aufwand dar. Eine Vorauswahl könnte ggf. mit Hilfe automatisierter Suche nach Stichworten innerhalb der Items oder der enthaltenen Konstruktbezeichnungen getroffen werden.

Darüber hinaus wurde in der vorliegenden Studie lediglich überprüft, inwiefern Inhaltsvalidierung durch die Verwendung von Panel-Daten möglich ist. Sind, wie im verwendeten Panel, Fragebogendaten enthalten, welche psychologische Konstrukte messen, so ist auch denkbar, Konstruktvalidität zu überprüfen. Hierfür müssten relevante Konstrukte identifiziert und entsprechende Korrelationen berechnet werden.

Auch die Überprüfung von Kriteriumsvalidität wäre mit einem Panel-Datensatz denkbar, wenn in den Daten ein relevantes Kriterium enthalten ist.

3.4.3 Fazit

Zusammenfassend zeigt die vorliegende Studie, dass die Verwendung von Panel-Daten in Kombination mit ML dafür geeignet ist, Ansätze für die Inhaltsvalidierung eines Fragebogens zu finden. Eine solche Inhaltsvalidierung kann nicht nur im Rahmen der Fragebogenentwicklung durchgeführt werden, sondern auch im Rahmen der Verbesserung und Evaluation bestehender Instrumente (Smith & McCarthy, 1995). Die in einer solchen Studie gewonnenen Erkenntnisse können schließlich als Anhaltspunkt dienen, an welchen Stellen die theoretische Konzeption des zu messenden Konstrukts evtl. nicht ganz klar ist bzw. in welchen Bereichen der Zusammenhang zu anderen Konstrukten im Rahmen konfirmatorischer Studien genauer untersucht werden sollte.

Kapitel 4

Allgemeine Diskussion

4.1 Zusammenfassung der Studien

In der vorliegenden Arbeit wurde der Einsatz von ML in der Fragebogenentwicklung exploriert. Dazu wurden zwei verschiedene Studien durchgeführt, welche an unterschiedlichen Aspekten der Fragebogenkonstruktion ansetzen.

In Studie I wurden Persönlichkeitseigenschaften basierend auf Bildbewertungen und Beantwortungszeiten vorhergesagt. Anschließend wurde untersucht, inwiefern aus den Vorhersagemodellen Messmodelle abgeleitet werden können. Die Ergebnisse dieser Studie zeigen, dass aus einem kombinierten Ansatz aus den im LASSO beibehaltenen Prädiktoren und theoriegeleitetem Ausschluss einzelner Prädiktoren, Messmodelle abgeleitet werden können. Die Sinnhaftigkeit dieser Modelle zur Messung des jeweiligen Konstrukts muss allerdings kritisch hinterfragt werden. Außerdem ist fraglich, ob diese Modelle auch auf andere Stichproben übertragbar sind. Inwiefern das gewählte Vorgehen Vorteile gegenüber traditionellen Methoden im Bereich der Fragebogenentwicklung aufweist, muss in weiteren Studien geklärt werden.

In Studie II wurden Persönlichkeitseigenschaften basierend auf Panel-Daten vorhergesagt und anschließend analysiert, inwiefern die Modellierungen ebenso wie Korrelationen zur Inhaltsvalidierung herangezogen werden können. Die Ergebnisse dieser Studie zeigen, dass die Prädiktoren mit den höchsten Regressionskoeffizienten Hinweise darauf liefern, welche Variablen mit dem Fragebogen zusammenhängen. Somit können sie als Ansatz zur Inhaltsvalidierung verwendet werden.

Aus den beiden beschriebenen Studien lassen sich zwei zentrale Ergebnisse ableiten:

1. ML kann in unterschiedlichen Bereichen der Fragebogenkonstruktion eingesetzt werden.
2. Lineare Modellierung scheint die betrachteten Zusammenhänge gut zu modellieren.

Im Folgenden werden die beiden zentralen Ergebnisse ausführlicher diskutiert.

4.1.1 ML in der Fragebogenkonstruktion

Die Ergebnisse beider Studien deuten darauf hin, dass ML in der Fragebogenentwicklung eingesetzt werden kann. Unabhängig davon, in welchem Schritt der Fragebogenkonstruktion ML eingesetzt wird, können die Ergebnisse verwendet werden, um Hinweise zu erhalten, welche Variablen mit dem zu messenden Konstrukt zusammenhängen. Diese Hinweise können auch genutzt werden, um Studien abzuleiten, welche die kausalen Zusammenhänge untersuchen und somit die theoretische Konzeption des Konstrukts ggf. zu erweitern (Shmueli & Koppius, 2011).

Im Folgenden wird auf die Vorteile der Verwendung von ML in verschiedenen Schritten der Fragebogenentwicklung eingegangen.

ML zur Erstellung von Messmodellen In Studie I wurde gezeigt, dass ML Algorithmen dazu verwendet werden können, um aus vielen verschiedenen Variablen einzelne zu bestimmen, welche in ein Messmodell für ein Kriterium integriert werden können. Klassischerweise werden zur Erstellung von Messmodellen explorative Faktorenanalysen herangezogen. Diese beiden Ansätze unterscheiden sich voneinander. Durch klassische Modellierung wird versucht, nachzuempfinden, welches Modell hinter der Erzeugung der Daten steckt (Breiman, 2001b). ML Algorithmen hingegen versuchen unabhängig von dem *wahren* Modell den Zusammenhang zwischen den Prädiktoren und dem Kriterium zu modellieren, sodass auch an unbekannten Daten möglichst gute Vorhersagen der Kriteriumsvariable gemacht werden können (Breiman, 2001b). Darüber hinaus wurden zur Vorhersage von Persönlichkeitseigenschaften verschiedene Methoden des überwachten Lernens verwendet. Dieser Ansatz entspricht eher einer konfirmatorischen Faktorenanalyse. Explorative Faktorenanalysen hingegen entsprechen einem Ansatz des unüberwachten Lernens, da diese Methode manifeste Variablen anhand ihrer geteilten Varianz Faktoren zuordnet, welche die gleiche latente Variable repräsentieren sollen (Goretzko et al., 2019).

Die Verwendung von ML zusammen mit Kreuzvalidierungsverfahren hat den Vorteil, dass die Vorhersagegüte der ML Modelle bereits an unbekannten Daten evaluiert wurde und somit ML Modelle ausgewählt werden können, welche auf unbekannte

Daten übertragbar sind (Breiman, 2001b; Shmueli, 2010). Inwiefern es durch Verwendung dieses Ansatzes möglich ist, die *wahre* Struktur zwischen den Items und der latenten Variable zu modellieren, muss in Simulationsstudien überprüft werden.

ML zur Validierung von Fragebögen In Studie II wurde gezeigt, dass ML im Rahmen des Fragebogenentwicklungsprozesses dazu verwendet werden kann, in umfangreichen Datensätzen Anhaltspunkte zur Inhaltsvalidität zu finden. Klassischerweise werden zur Überprüfung der Validität von Fragebögen eigene Studien entwickelt, in denen ausgewählte Variablen erhoben werden (Bleidorn et al., 2017). Diese Studien sind sehr aufwendig und können nur eine begrenzte Anzahl an Variablen enthalten. Daher scheint es sinnvoll, wenn möglich, auf große Datensätze zurückzugreifen, in denen verschiedene Variablen enthalten sind. Dies kann beispielsweise durch die Integration des zu validierenden Fragebogens in eine Panel-Befragung umgesetzt werden.

In Studie II zeigte das LASSO eine gute Vorhersagegüte. Die durch diesen Algorithmus entstandenen Modelle haben gegenüber normaler Korrelationsanalysen zwei Vorteile. Einerseits wird durch die Regularisierung lediglich ein Teil der vorhandenen Prädiktoren im Modell beibehalten, sodass gerade bei großen Datensätzen nur ein Teil der Prädiktoren betrachtet werden muss. Andererseits werden die Prädiktoren in Abhängigkeit zueinander modelliert. Weisen Variablen hohe Korrelationen zueinander auf, so wird im LASSO lediglich eine dieser Variable beibehalten (Zou & Hastie, 2005). Dies kann jedoch auch den Nachteil haben, dass hypothetisierte Zusammenhänge nicht gefunden werden, da sie hohe Korrelationen mit anderen Variablen aufweisen und daher nicht im Modell beibehalten wurden. Alternative Methoden, welche dieses Problem umgehen, werden in Kapitel 4.1.2 diskutiert.

Allgemeine methodische Vorschläge In beiden Studien der vorliegenden Arbeit wurden alle Prädiktoren auf einmal integriert. Ein vielversprechender Ansatz zur Überprüfung der Wichtigkeit von Variablen, welche entweder geteilte Eigenschaften haben oder inhaltliche Ähnlichkeit aufweisen, scheint der von Yarkoni und Westfall (2017) beschriebene zu sein. Diese schlagen vor, innerhalb eines ML Algorithmus die Vorhersagegüte eines Modells mit allen Prädiktoren mit Modellen zu vergleichen, in denen lediglich ein Teil der gegebenen Variablen enthalten ist (Yarkoni & Westfall, 2017). Die Autoren stellen heraus, dass auf diese Weise herausgefunden werden kann, inwiefern Prädiktoren-Gruppen einen unterschiedlich starken Einfluss auf die Vorhersagegüte haben (Yarkoni & Westfall, 2017).

Übertragen auf Studie I hätten somit die Variablen der drei verschiedenen Kategorien *Bewertung, Beantwortungszeit* und *Interaktionsterm* einzeln in die Vorher-

sagen integriert werden können und somit die Frage beantwortet, welche der drei Informationsquellen den größten Einfluss auf die Vorhersage hat.

Übertragen auf Studie II wäre es möglich, Variablen anhand ihrer Beschreibung inhaltlich zu gruppieren. Hierzu wäre es in einem ersten Schritt notwendig, die Beschreibungen der Variablen aus dem Handbuch zu extrahieren und über einen sprachbasierten Algorithmus in inhaltliche Cluster zu unterteilen. Alternativ kann dies manuell geschehen. Anschließend könnten die Variablen-Gruppen einzeln integriert werden. Gerade in Bezug auf Inhaltsvalidierung wäre dieses Vorgehen informativ, da so genauere Informationen über die Wichtigkeit einzelner inhaltlicher Themen zur Vorhersage des Kriteriums gewonnen werden könnten.

Einen alternativen Ansatz, um in der psychologischen Forschung mit großen Datensätzen umzugehen, schlagen Cheung und Jak (2016) vor. Dieser scheint auch für den Einsatz von ML in der Fragebogenentwicklung anwendbar. Cheung und Jak (2016) empfehlen, den Datensatz in mehrere Teildatensätze aufzuteilen und die Ergebnisse von Analysen in den Teildatensätzen anschließend mit Hilfe metaanalytischer Verfahren zu integrieren. Um einen solchen Ansatz für die Fragebogenentwicklung zu verwenden, könnten beispielsweise in einem Teildatensatz durch eine explorative Faktorenanalyse die Zuordnungen der verschiedenen Variablen zu Faktoren bestimmt werden. Anschließend könnte diese Faktorstruktur in eine konfirmatorische Faktorenanalyse übertragen werden, welche an anderen Teildatensätzen überprüft wird. Somit könnten einerseits Messmodelle auf Basis der integrierten Prädiktoren erstellt werden. Andererseits könnten Variablen identifiziert werden, welche auf die gleichen Faktoren laden wie die Items des zu betrachtenden Konstrukts. Da dies ebenfalls Hinweise darauf gibt, zu welchen anderen Items die zu betrachtenden Items ähnlich sind (Goretzko et al., 2019), könnte so ebenfalls eine Inhaltsvalidierung erfolgen.

4.1.2 Lineare Modellierung

In beiden Studien haben LASSO Modellierungen, im Vergleich zu anderen ML Algorithmen, eine gute Vorhersagegüte erreicht. Dies deutet darauf hin, dass lineare Modelle die zugrundeliegenden Zusammenhänge zwischen den Variablen gut annähern können.

Im Rahmen der Fragebogenentwicklung sind Fragestellungen relevant, welche eine hohe Interpretierbarkeit der entstehenden Modelle erfordern. Daher scheint es sinnvoll, in diesem Rahmen gut interpretierbare Modelle vorzuziehen, welche ggf. eine geringere Vorhersagegüte aufweisen als komplexere Modelle (Shmueli & Koppius,

2011).

Modellierung von Interaktionen im LASSO In den Studien der vorliegenden Arbeit wurde jeweils eine reine Form des LASSO verwendet. Zukünftig sollte überprüft werden, inwiefern Weiterentwicklungen dieser Methode, welche immer noch gut interpretierbare Modelle bieten, aber dennoch in der Lage sind, auch komplexere Zusammenhänge abzubilden, Vorteile mit sich bringen.

Beispielsweise könnten Weiterentwicklungen des LASSO verwendet werden, welche in der Lage sind, Interaktionen zwischen den Prädiktoren zu modellieren. Dies könnte durch ein Boosted LASSO (BLASSO) (Zhao & Yu, 2004) oder ein hierarchisches Group-LASSO (glinternet, group-LASSO interaction network) (Lim & Hastie, 2015) realisiert werden.

Die Variablenselektion im *BLASSO* findet ähnlich wie bei der Erstellung eines Entscheidungsbaums in Fitting-Schritten statt (Zhao & Yu, 2004). Hierbei findet in jedem Schritt zunächst eine Vorwärtsselektion statt, welche analog zu Boosting und Forward Stagewise Fitting funktioniert (Zhao & Yu, 2004). Im Entscheidungsbaum entspricht dies dem *growing*. Anschließend wird eine Rückwärtsselektion durchgeführt, wobei ein LASSO verwendet wird, um Fehler aus vorherigen Schritten rückgängig zu machen (Zhao & Yu, 2004). Im Entscheidungsbaum entspricht dies dem *pruning*. Somit wird jeder Fitting-Schritt als Teilproblem der Optimierung gesehen (Zhao & Yu, 2004). Dabei können im BLASSO unterschiedliche Verlustfunktionen und verschiedene Modellierungen des zugrunde liegenden Learners (Base-Learner) verwendet werden (Zhao & Yu, 2004).

Das *glinternet* hingegen verfolgt einen hierarchischen Ansatz, bei welchem zunächst Haupteffekte und anschließend Interaktionen integriert werden. Die Variablenselektion wird durch die Verwendung eines Gruppen-LASSO (Yuan & Lin, 2006) durchgeführt (Lim & Hastie, 2015). Da es sich um ein streng hierarchisches Verfahren handelt, werden lediglich Interaktionen im Modell beibehalten, wenn auch die beiden dazugehörigen Haupteffekte im Modell beibehalten werden (Lim & Hastie, 2015). Da diese Methode sowohl auf kategoriale als auch numerische Variablen und eine Kombination dieser angewendet werden kann (Lim & Hastie, 2015), scheint sie bei der Verwendung psychologischer Daten gut einsetzbar.

Die beiden zuvor beschriebenen Verfahren sollten demnach verwendet werden, wenn angenommen wird, dass Interaktionseffekte zwischen den Prädiktoren bestehen.

Alternative regularisierte Regressionen Wie oben beschrieben, hat das LASSO den Nachteil, dass bei korrelierten Variablen lediglich eine zufällig ausgewählt

und im Modell beibehalten wird (Zou & Hastie, 2005). Darüber hinaus können sie aufgrund ihres Strafterms im Falle p>n[1] maximal n Prädiktoren im Modell beibehalten (Zou & Hastie, 2005). Zudem konnte gezeigt werden, dass Ridge Regression im Fall von p>n mit korrelierten Prädiktoren eine bessere Vorhersagegüte als das LASSO aufweist (Zou & Hastie, 2005).

Vor allem bei der Verwendung von Panel-Daten scheint es nicht ausgeschlossen, dass sowohl eine p>n-Situation auftritt als auch korrelierte Prädiktoren vorliegen. Daher sollten in zukünftiger Forschung vermehrt lineare Methoden miteinander verglichen werden, welche unterschiedliche Eigenschaften aufweisen.

Methoden, welche nicht die zuvor beschriebenen Nachteile des LASSO aufweisen sind beispielsweise Elastic Net (Zou & Hastie, 2005), adaptive LASSO (Zou, 2006), relaxed LASSO (Meinshausen, 2007) und das Random LASSO (Wang, Nan, Rosset & Zhu, 2011). Von diesen Methoden werden das Elastic Net und das adaptive LASSO am häufigsten eingesetzt[2], weshalb im Folgenden lediglich diese beiden Ansätze beschrieben werden.

Im Elastic Net wird analog zum LASSO eine Variablenselektion durch die Einführung eines Strafterms durchgeführt (Zou & Hastie, 2005). Im Gegensatz zum LASSO ist der Strafterm so angelegt, das eine Variablenselektion durchgeführt werden kann, aber auch alle Prädiktoren im Modell beibehalten werden können (Zou & Hastie, 2005). Dadurch kann das Elastic Net als eine Generalisierung des LASSO angesehen werden(Zou & Hastie, 2005). Darüber hinaus behält das Elastic Net im Falle von wichtigen, korrelierten Variablen all diese Variablen im Modell und wählt nicht wie das LASSO eine für das Modell aus (Zou & Hastie, 2005). Diese Methode sollte demnach verwendet werden, wenn es gewünscht ist, im Falle von wichtigen, korrelierten Prädiktoren alle beizubehalten oder eine p>n-Situation vorliegt.

Das adaptive LASSO stellt eine Weiterentwicklung des LASSO dar, in welchem die einzelnen Koeffizienten individuell regularisiert werden (Zou, 2006). Dies ermöglicht der Methode, ein konsistentes Set an Prädiktoren im Modell zu behalten, welches asymptotisch dem wahren Modell entspricht und gleichzeitig eine optimale Vorhersagegüte zu erreichen (Zou, 2006). Diese Methode sollte demnach angewendet werden, wenn es wichtig ist, eine konsistente Auswahl an wichtigen Prädiktoren zu erhalten.

Die zuvor beschriebenen Methoden stellen lediglich einen Teil der linearen Methoden dar, welche in zukünftiger Forschung berücksichtigt werden könnten.

---

[1] p = Anzahl der Prädiktoren, n = Anzahl der Beobachtungen

[2] Diese Methoden weisen mit Abstand die meisten Zitationen auf.

4.2 Fazit

Der Einsatz von ML in der Fragebogenentwicklung diente der Exploration dessen, inwiefern moderne Methoden zur Beantwortung traditioneller psychologischer Fragestellungen eingesetzt werden können. Die Ergebnisse der vorliegenden Arbeit deuten darauf hin, dass ML in der Fragebogenkonstruktion in unterschiedlichen Schritten des Entwicklungsprozesses eingesetzt werden können. Unabhängig von der Zielsetzung zeigten in der vorliegenden Arbeit regularisierte lineare Modelle eine hohe Vorhersagegüte, weshalb angenommen werden kann, dass diese die Zusammenhänge gut annähern können. Obwohl regularisierte lineare Modelle eine gute Interpretierbarkeit aufweisen, wurde deutlich, dass diese Methoden vor allem explorative Fragestellungen beantworten können. Hierbei können wertvolle Informationen über den betrachteten Fragebogen gewonnen werden. Dies bedeutet gleichzeitig, dass sie durch klassische Verfahren wie beispielsweise konfirmatorische Analysen komplementiert werden sollten.

Dementsprechend kann geschlussfolgert werden, dass Moderne und Tradition zumindest in der psychologischen Fragebogenentwicklung Synergien aufweisen.

Anhang A

Studie I

A.1 Deskriptive Statistiken

Tabelle A.1

Deskriptive Statistiken der Bildbewertungen

	M	*SD*	*Mdn*	Min	*Max*	Schiefe	Kurtosis	*SE*
O1	3.32	1.13	3	1	5	-0.18	-0.85	0.05
O2	3.38	1.26	4	1	5	-0.28	-1.04	0.05
O3	3.08	1.16	3	1	5	-0.07	-0.87	0.05
O4	3.30	1.03	3	1	5	-0.31	-0.58	0.04
O5	2.85	1.23	3	1	5	0.17	-0.97	0.05
O6	3.68	0.99	4	1	5	-0.46	-0.38	0.04
O7	3.63	1.07	4	1	5	-0.50	-0.47	0.05
O8	2.92	1.36	3	1	5	0.12	-1.25	0.06
O9	2.68	1.10	3	1	5	0.17	-0.78	0.05
O10	3.62	1.05	4	1	5	-0.44	-0.53	0.05
G1	3.01	1.18	3	1	5	-0.06	-0.81	0.05
G2	2.91	1.14	3	1	5	0.00	-0.88	0.05
G3	4.53	0.76	5	1	5	-1.75	2.79	0.03
G4	3.73	1.18	4	1	5	-0.72	-0.36	0.05
G5	2.16	1.01	2	1	5	0.59	-0.32	0.04
G6	3.82	0.93	4	1	5	-0.65	0.14	0.04
G7	2.28	1.04	2	1	5	0.50	-0.51	0.05
G8	2.80	1.11	3	1	5	0.14	-0.79	0.05
G9	3.05	1.14	3	1	5	-0.23	-0.79	0.05
G10	3.09	1.09	3	1	5	-0.04	-0.75	0.05
E1	3.38	1.03	3	1	5	-0.19	-0.67	0.04
E2	3.82	0.92	4	1	5	-0.41	-0.51	0.04
E3	1.82	0.89	2	1	5	1.06	0.97	0.04
E4	3.92	0.93	4	1	5	-0.90	0.78	0.04
E5	3.38	1.17	4	1	5	-0.32	-0.81	0.05
E6	3.57	1.05	4	1	5	-0.47	-0.32	0.05
E7	2.39	1.20	2	1	5	0.46	-0.78	0.05
E8	2.46	0.99	2	1	5	0.23	-0.68	0.04
E9	2.98	1.09	3	1	5	-0.12	-0.82	0.05
E10	4.04	0.91	4	1	5	-0.91	0.52	0.04

Anmerkungen. M = Mittelwert; SD = Standardabweichung; Md = Median; Min = Minimum, Max = Maximum, SE = Standardfehler. O1-O10 Bilder zu Offenheit, G1-G10 Bilder zu Gewissenhaftigkeit, E1-E10 Bilder zu Extraversion. N= 530. Alle dargestellten Variablen wurden auf einer 5-stufigen Likert-basierten Skala bewertet.

Tabelle A.2

Deskriptive Statistiken der Bearbeitungszeit pro Bildbewertung

	M	SD	Md	Min	Max	Schiefe	Kurtosis	SE
ZO1	0.86	6.20	0.00	-15.00	105.00	9.98	151.77	0.27
ZO2	1.42	4.75	0.00	-7.00	50.00	4.51	33.34	0.21
ZO3	0.24	6.35	-0.50	-15.00	75.00	7.27	69.68	0.28
ZO4	1.90	6.87	0.00	-9.00	89.00	5.73	53.90	0.30
ZO5	2.05	5.52	1.00	-21.00	44.50	3.38	19.34	0.24
ZO6	0.14	7.00	-1.00	-31.00	86.00	6.33	60.24	0.30
ZO7	0.25	6.49	-1.00	-7.00	76.00	7.26	66.01	0.28
ZO8	1.46	6.00	0.00	-12.00	68.00	6.03	52.99	0.26
ZO9	-0.12	5.09	-1.00	-18.50	50.00	5.13	38.22	0.22
ZO10	0.52	6.22	0.00	-16.00	89.00	7.81	90.06	0.27
ZG1	4.16	8.34	2.00	-10.00	95.50	5.20	42.15	0.36
ZG2	1.39	5.58	0.00	-6.50	49.00	4.87	30.72	0.24
ZG3	0.30	5.72	0.00	-16.00	62.00	5.98	50.08	0.25
ZG4	3.33	7.84	1.50	-11.00	88.00	5.99	48.43	0.34
ZG5	0.48	7.05	-0.50	-9.00	85.50	7.95	77.81	0.31
ZG6	1.10	6.58	0.00	-10.00	102.50	8.60	111.99	0.29
ZG7	0.09	5.14	-1.00	-9.50	80.50	8.83	118.59	0.22
ZG8	1.60	6.85	0.00	-17.00	77.00	6.27	54.32	0.30
ZG9	1.14	5.57	0.00	-11.50	64.00	4.97	41.27	0.24
ZG10	-0.11	4.68	-0.50	-24.50	83.50	10.66	191.27	0.20
ZE1	1.70	7.03	0.00	-12.50	84.00	6.99	66.15	0.31
ZE2	2.88	6.72	1.00	-8.50	65.00	4.49	28.23	0.29
ZE3	0.77	7.49	0.00	-40.00	79.50	5.72	48.17	0.33
ZE4	1.21	4.99	0.00	-9.00	48.50	4.81	35.47	0.22
ZE5	2.56	8.85	1.00	-43.00	103.00	6.22	57.89	0.38
ZE6	0.87	4.57	0.00	-41.00	48.50	2.21	39.33	0.20
ZE7	2.99	8.88	1.00	-31.00	102.50	5.90	51.74	0.39
ZE8	1.21	7.22	0.00	-10.00	92.00	7.97	81.37	0.31
ZE9	1.83	6.29	0.00	-14.00	68.00	5.51	43.17	0.27
ZE10	-0.67	5.56	-1.00	-11.50	77.00	8.08	92.59	0.24

Anmerkungen. M = Mittelwert; SD = Standardabweichung; Md = Median; Min = Minimum, Max = Maximum, SE = Standardfehler. ZO1-ZO10 Bearbeitungszeit der Bilder zu Offenheit, ZG1-ZG10 Bearbeitungszeit der Bilder zu Gewissenhaftigkeit, ZE1-ZE10 Bearbeitungszeit der Bilder zu Extraversion. N= 530. Alle dargestellten Variablen stellen die absolute Abweichung zum Median der Antwortzeit pro Person dar. Ein positiver Wert bedeutet, dass die Teilnehmer zur Beantwortung des Bildes X Sekunden länger als im Median benötigt haben. Eine negative Zahl bedeutet, dass die Teilnehmer zur Beantwortung des Bildes X Sekunden kürzer als im Median benötigt haben.

Tabelle A.3

Deskriptive Statistiken der BFI-S Skalen

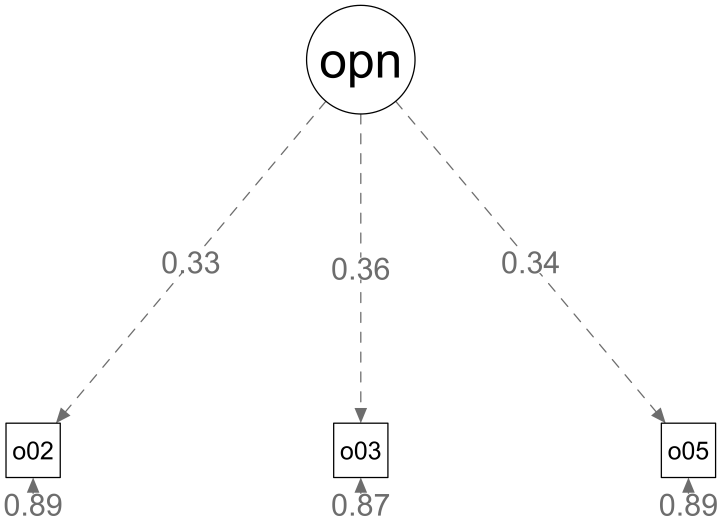
	M	SD	Md	Min	Max	Schiefe	Kurtosis	SE
Offenheit	5.02	1.15	5.17	2	7	-0.40	-0.45	0.05
Gewissenhaftigkeit	5.23	1.03	5.33	2	7	-0.38	-0.31	0.04
Extraversion	4.70	1.34	4.67	1	7	-0.28	-0.57	0.06
Verträglichkeit	5.38	1.00	5.33	2	7	-0.64	0.20	0.04
Neurotizismus	4.17	1.32	4.00	1	7	-0.02	-0.63	0.06

Anmerkungen. M = Mittelwert; SD = Standardabweichung; Md = Median; Min = Minimum, Max = Maximum, SE = Standardfehler. N= 530. Alle dargestellten Variablen wurden auf einer 5-stufigen Likert-basierten Skala bewertet. Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit, Neurotizismus = jeweiliger Skalenmittelwert auf Basis der BFI-S Items.

A.2 CFA

Abbildung A.1

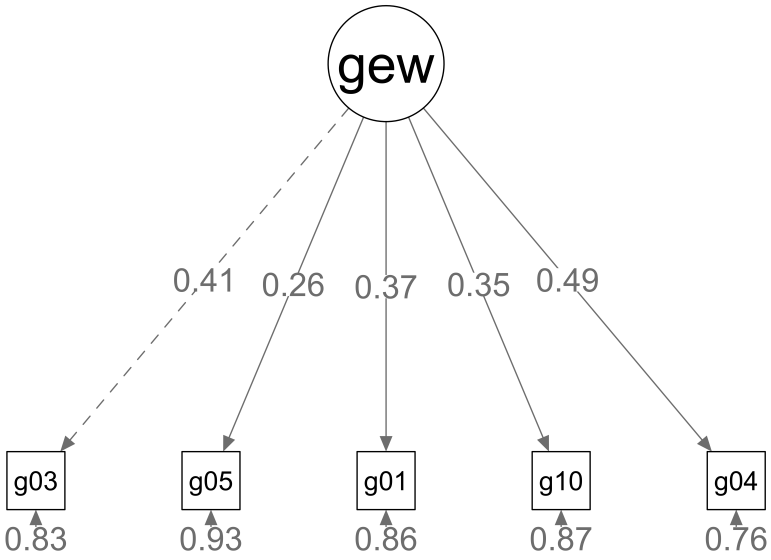
Strukturgleichungsmodell der CFA zur Modellierung von Offenheit



Es werden standardisierte Ladungen dargestellt. Gestrichelte Linien zeigen an, dass die Ladung auf 1 fixiert wurde. $\chi^2(2) = 2.192$, $p = 0.334$, p (Bollen-Stine Bootstrap) = 0.369, CFI = 0.990, RMSEA = 0.013, SRMR = 0.026.

Abbildung A.2

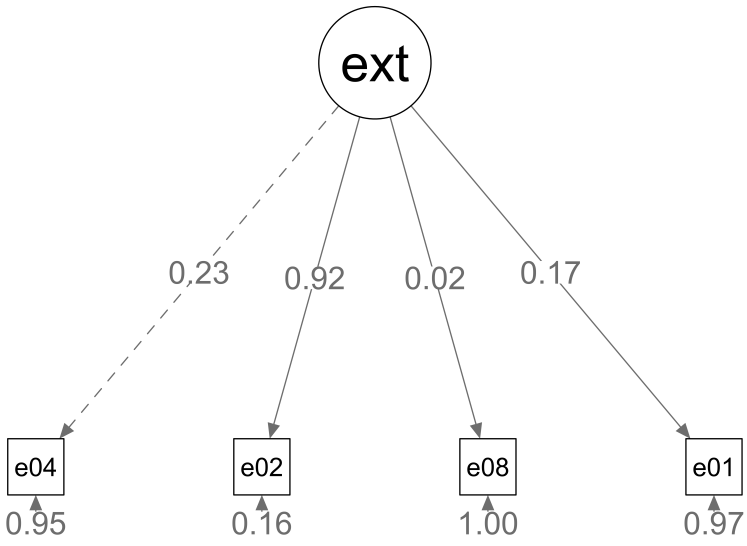
Strukturgleichungsmodell der CFA zur Modellierung von Gewissenhaftigkeit



Es werden standardisierte Ladungen dargestellt. Gestrichelte Linien zeigen an, dass die Ladung auf 1 fixiert wurde. $\chi^2(9) = 22.078$, $p = 0.009$, p (Bollen-Stine Bootstrap) = 0.015, CFI = 0.857, RMSEA = 0.052, SRMR = 0.053.

Abbildung A.3

Strukturgleichungsmodell der CFA zur Modellierung von Extraversion



Es werden standardisierte Ladungen dargestellt. Gestrichelte Linien zeigen an, dass die Ladung auf 1 fixiert wurde. $\chi^2(2) = 1.066$, $p = 0.587$, p (Bollen-Stine Bootstrap) = 0.590, CFI = 1.000, RMSEA = 0.000, SRMR = 0.014.

A.3 EFA

Um die Ergebnisse der Modellierungen auf Basis der Vorhersageergebnisse der ML Algorithmen mit den Ergebnissen zu vergleichen, welche klassische Methoden liefern, wurden explorative Faktorenanalysen gerechnet.

Zunächst wurde eine EFA gerechnet, in welcher alle Bildbewertungen inkludiert wurden. Da die ausgewählten Bilder die drei Persönlichkeitseigenschaften Offenheit, Gewissenhaftigkeit und Extraversion darstellen sollten, wurden drei Faktoren angenommen. Die Faktorladungen sind Tabelle A.4 zu entnehmen.

Da die Zusammenhänge zwischen den Prädiktoren und Kriterien univariat modelliert wurden und die abgeleiteten Messmodelle daher lediglich eine Persönlichkeitseigenschaft abbilden, wurden darüber hinaus EFAs gerechnet, in denen lediglich die Bewertungen einer Persönlichkeitseigenschaft enthalten waren. Die Anzahl der Faktoren wurde hierfür durch eine Parallel-Analyse nach Horn (1965) bestimmt.

Auf Basis der Ergebnisse der Parallel-Analyse wurden für Offenheit drei Faktoren angenommen. Die Faktorladungen der entsprechenden EFA sind Tabelle A.5 zu entnehmen.

Die Parallel-Analyse für Gewissenhaftigkeit nahm null Faktoren an, weshalb keine EFA hierfür gerechnet wurde.

Auf Basis der Ergebnisse der Parallel-Analyse wurden für Extraversion vier Faktoren angenommen. Die Faktorladungen der entsprechenden EFA sind Tabelle A.6 zu entnehmen.

Tabelle A.4

Faktorladungen der explorativen Faktorenanalyse mit allen Bildbewertungen

	MR1	MR3	MR2
O1	**-0.46**	0.17	-0.04
O2	-0.31	0.03	**0.37**
O3	0.16	0.04	**0.50**
O4	0.02	0.12	-0.17
O5	-0.18	0.00	0.06
O6	-0.18	**0.54**	-0.09
O7	-0.04	0.24	0.12
O8	-0.31	-0.06	**0.42**
O9	0.25	0.04	**0.35**
O10	-0.01	0.09	0.30
G1	0.21	0.07	0.07
G2	-0.05	0.31	0.07
G3	0.22	0.34	-0.09
G4	**0.45**	0.29	-0.08
G5	**0.51**	-0.02	-0.00
G6	0.04	**0.37**	0.07
G7	0.29	-0.17	-0.08
G8	0.27	0.00	**0.36**
G9	0.10	0.24	0.05
G10	**0.35**	0.19	0.19
E1	-0.26	0.32	-0.03
E2	-0.29	0.17	-0.19
E3	0.13	-0.19	-0.21
E4	-0.00	0.18	-0.25
E5	0.34	0.07	-0.04
E6	-0.27	0.06	0.04
E7	-0.16	0.03	-0.03
E8	0.18	-0.07	-0.27
E9	-0.02	-0.21	-0.27
E10	-0.10	**0.60**	-0.23

Anmerkung. Faktorladungen $\geq |0.35|$ sind fett hervorgehoben.

Tabelle A.5

Faktorladungen der explorativen Faktorenanalyse mit den Bewertungen für Offenheit

	MR1	MR2	MR3
O1	0.12	-0.10	**0.37**
O2	**0.61**	0.03	0.14
O3	-0.06	**0.98**	-0.13
O4	-0.07	-0.05	-0.06
O5	0.02	0.05	0.24
O6	0.01	-0.08	**0.54**
O7	-0.09	0.12	0.20
O8	**0.90**	-0.08	-0.12
O9	0.08	0.25	-0.26
O10	0.01	0.32	0.09

Anmerkung. Faktorladungen $\geq |0.35|$ sind fett hervorgehoben.

Tabelle A.6

Faktorladungen der explorativen Faktorenanalyse mit den Bewertungen für Extraversion

	MR2	MR4	MR1	MR3
E1	-0.02	0.08	-0.01	**0.37**
E2	0.06	0.20	**0.45**	0.24
E3	**0.62**	-0.05	-0.05	-0.03
E4	0.13	-0.04	0.30	0.11
E5	0.08	0.06	-0.25	0.02
E6	-0.17	0.25	**0.38**	-0.06
E7	0.06	**0.82**	-0.07	-0.05
E8	**0.60**	0.10	-0.20	0.09
E9	**0.38**	-0.01	0.25	-0.20
E10	-0.03	-0.17	0.12	**0.49**

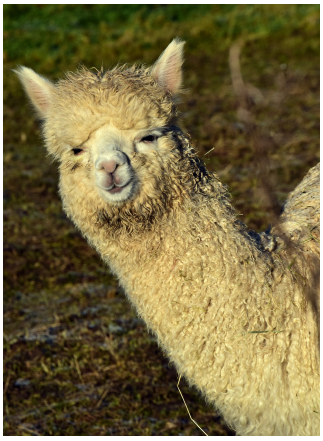
Anmerkung. Faktorladungen $\geq |0.35|$ sind fett hervorgehoben.

A.4 Verwendete Bilder

Abbildung A.4

Bilder

(a) *O1*



(b) *O2*

(c) *O3*

(d) *O4*

(e) *O5*

(f) *O6*

Quellen:
O1: pixabay.com/de/alpaka-s%C3%A4ugetier-tier-pelz-natur-3023311
O2: pixabay.com/de/buch-traum-reise-fantasie-2899636
O3: pixabay.com/de/bustos-filsofia-aristoteles-756620
O4: pixabay.com/de/kinder-jungen-fechten-vorschau-958474
O5: pixabay.com/de/farben-color-bunt-kreativ-malen-2413936
O6: pixabay.com/de/pusteblume-l%C3%B6wenzahn-himmel-blume-463928

Abbildung A.5

Bilder (Fortsetzung 1)

(a) *O7*

(b) *O8*

(c) *O9*

(d) *O10*

(e) *G1*

(f) *G2*

Quellen:
O7: pixabay.com/de/w%C3%BCste-sand-landschaft-sonne-790640
O8: pixabay.com/de/fantasie-schildkr%C3%B6te-taube-wolke-3105819
O9: pexels.com/photo/white-black-game-fun-33078
O10: pixabay.com/de/wissen-buch-bibliothek-gl%C3%A4ser-1052011
G1: pexels.com/photo/coffee-cup-mug-desk-5251
G2: pixabay.com/de/insekt-lebensmittel-biene-honig-3155574

Abbildung A.6

Bilder (Fortsetzung 2)

(a) *G3*

(b) *G4*

(c) *G5*

(d) *G6*

(e) *G7*

(f) *G8*

Quellen:
G3: pixabay.com/de/m%C3%BCll-m%C3%BCllcontainer-abfall-2729608
G4: pixabay.com/en/mistake-spill-slip-up-accident-876597
G5: pixabay.com/en/koala-bear-australia-teddy-sleep-9960
G6: pixabay.com/de/baum-natur-herbst-landschaft-holz-3094059
G7: pixabay.com/de/die-dschungel-von-chiapas-1865639
G8: pixabay.com/de/uhrmacher-vietnam-asien-saigon-1163433

Abbildung A.7

Bilder (Fortsetzung 3)

(a) *G9*

(b) *G10*

(c) *E1*

(d) *E2*

(e) *E3*

(f) *E4*

Quellen:
G9: pixabay.com/de/m%C3%BCll trennung-m%C3%BCll tonnen-recycling-502952
G10: pixabay.com/de/pfeil-ziel-bullseye-kreis-zentrum-2886223
E1: pixabay.com/de/erdm%C3%A4nnchen-familie-zoo-tiere-2235119
E2: pexels.com/de/foto/baume-chat-draussen-frauen-745045
E3: pixabay.com/de/meditation-berge-nachdenklich-3000172
E4: pixabay.com/de/restaurant-menschen-essen-690975

Abbildung A.8

Bilder (Fortsetzung 4)

(a) *E5*

(b) *E6*

(c) *E7*

(d) *E8*

(e) *E9*

(f) *E10*

Quellen:
E5: pixabay.com/de/ja-nein-m%C3%B6glichkeit-entscheidung-3100993
E6: pexels.com/de/foto/abenteuer-aktivitat-backpacking-baume-344100
E7: pexels.com/de/foto/abenteuer-berge-bildung-fallschirmspringen-70361
E8: pixabay.com/en/boy-aloof-kid-alone-thinking-2671425
E9: pixabay.com/de/kampagne-haus-prato-green-natur-399672
E10: pixabay.com/de/g%C3%A4nsebl%C3%BCmchen-blume-bl%C3%BCte-3102518

Anhang B

Studie II

Tabelle B.1

Deskriptive Statistiken des BFI-S

	N	*M*	*SD*	*Mdn*	*Min*	*Max*	Schiefe	Kurtosis
Offenheit	2964	3.34	0.86	3.50	1	5	-0.11	-0.52
Gewissenhaftigkeit	2968	3.91	0.71	4.00	1	5	-0.35	-0.22
Extraversion	2969	3.14	0.86	3.00	1	5	0.01	-0.59
Verträglichkeit	2966	3.16	0.70	3.00	1	5	-0.14	-0.25
Neurotizismus	2967	2.81	0.81	3.00	1	5	0.16	-0.42

Literatur

Abdi, H. & Williams, L. J. (2010). Principal components analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. doi:10.1002/wics.101

Ahmad, N. & Siddique, J. (2017). Personality assessment using Twitter tweets. *Procedia Computer Science*, *112*, 1964–1973. doi:10.1016/j.procs.2017.08.067

Alexander, D. L. J., Tropsha, A. & Winkler, D. A. (2015). Beware of R2: Simple, unambigous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling*, *55*(7), 1316–1322. doi:10.1021/acs.jcim.5b00206

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Survey*, *4*, 40–79. doi:10.1214/09-SS054

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602–614. doi:10.1037/0003-066X.63.7.602

Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R. & Sinha, V. (2017). 25 Tweets to know you: A new model to predict personality with social media. In *Proceeding of the 11th International AAAI Conference on Web and Social Media* (S. 472–475).

Azucar, D., Marengo, D. & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. doi:10.1016/j.paid.2017.12.018

Bai, S., Zhu, T. & Cheng, L. (2012). Big-Five personality prediction based on user behaviors at social network sites. arXiv Preprint: 1204.4809

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi:10.1177/1745691612459060

Barrick, M. R. & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Behrend, T. S., Sharek, D. J., Meade, A. W. & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. doi:10.3758/s13428-011-0081-0

Binder, M. (2019). mlrCPO: Composable Preprocessing Operators and Pipelines for Machine Learning.

Bischl, B. & Lang, M. (2015). parallelMap: Unified Interface to parallelization backends.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research*, *17*(1), 1–5.

Bleidorn, W., Hopwood, C. J. & Wright, A. G. (2017). Using big data to advance personality theory. *Current Opinion in Behavioral Sciences*, *18*, 79–82. doi:10.1016/j.cobeha.2017.08.004

Bleidorn, W. & Hopwood, C. J. (2019). Using Machine Learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. doi:10.1177/1088868318772990

Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061

Boyd, R. L. & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. doi:10.1016/j.cobeha.2017.07.017

Boyle, G. J., Matthews, G. & Saklofske, D. H. (2008). Personality Measurement and Testing: An Overview. In G. J. Boyle, G. Matthews & D. H. Saklofske (Hrsg.), *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing* (Bd. 2, S. 1–26). doi:10.4135/9781849200479

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. doi:10.1214/ss/1009213726

Brick, C. & Lewis, G. J. (2016). Unearthing the "green" personality: Core traits predict environmentally friendly behavior. *Environment and Behavior*, *48*(5), 635–658. doi:10.1177/0013916514554695

Buhrmester, M., Kwang, T. & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Butrus, N. & Witenberg, R. T. (2013). Some personality predictors of tolerance to human diversity: The roles of openness, agreeableness, and empathy. *Australian Psychologist*, *48*(4), 290–298. doi:10.1111/j.1742-9544.2012.00081.x

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. doi:10.1037/h0046016

Cattell, R. B. (1946). *Description and measurement of personality*. New York: World Book Company.

Celli, F. (2012). Unsupervised personality recognition for social network sites. In *Proceedings of the International Conference on Digital Society* (S. 59–62).

Chavent, M., Kuentz-Simonet, V., Liquet, B. & Saracco, L. (2012). ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software*, *50*, 1–16. doi:10.18637/jss.v050.i13

Chavent, M., Kuentz-Simonet, V., Labenne, A. & Saracco, J. (2017). Multivariate Analysis of Mixed Data: The R Package PCAmixdata. arXiv Preprint: 1411.4911

Cheung, F. M., van de Vijver, F. J. & Leong, F. T. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, *66*(7), 593–603. doi:10.1037/a0022389

Cheung, M. W. & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, *7*, 738. doi:10.3389/fpsyg.2016.00738

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. doi:10.1037//1040-3590.7.3.309

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Corfield, D. & Williamson, J. (Hrsg.). (2013). *Foundations of Bayesianism*. Dordrecht: Springer Science & Business Media.

Costa, P. T. & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology*, *54*(5), 853–863. doi:10.1037/0022-3514.54.5.853

Cristani, M., Vinciarelli, A., Segalin, C. & Perina, A. (2015). Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *21st ACM international conference on Multimedia* (S. 21–25). doi:10.1145/2502081.2502280

Cronbach, P. E. & Meehl, L. J. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi:10.1037/h0040957

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A. & Swinton, J. (2018). xtable: Export tables to LaTeX or HTML.

De Carolis, B. & Mazzotta, I. (2011). Motivating people in smart environments. In L. Ardissono & T. Kuflik (Hrsg.), *Advances in User Modeling. UMAP 2011. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer.

De Raad, B., Perugini, M., Hrebickova, M. & Szarota, P. (1998). Lingua Franca of personality: Taxonomies and structures based on the psycholexical approach. *Journal of Cross-Cultural Psychology*, *29*(1), 212–232. doi:10.1177/0022022198291011

de Bruijne, M. & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review*, *31*(4), 482–504. doi:10.1177/0894439313483976

DeNeve, K. M. & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin*, *124*(2), 197–229. doi:10.1037/0033-2909.124.2.197

Döring, A. K., Blauensteiner, A., Aryus, K., Drögekamp, L. & Bilsky, W. (2010). Assessing values at an early age: The Picture-Based Value Survey for Children (PBVS-C). *Journal of Personality Assessment*, *92*(5), 439–448. doi:10.1080/00223891.2010.497423

Dougherty, E. R., Kim, S. & Chen, Y. (2000). Coefficient of determination in nonlinear signal processing. *Signal Processing*, *80*(10), 2219–2235. doi:10.1016/S0165-1684(00)00079-7

Dowle, M. & Srinivasan, A. (2018). data.table: Extension of 'data.frame'.

Drucker, H., Burges, C. J. C., Kaufmann, L., Smola, A. & Vapnik, V. (1996). Support vector regression machines. In *Advances in Neural Information Processing Systems* (S. 155–161).

Dunn, T. G., Lushene, R. E. & O'Neil, H. F. (1972). Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology*, *39*(3), 381–387. doi:10.1037/h0033855

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F. & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, *27*, 6137–6157. doi:10.1002/sim.3429

Eftekhar, A., Fullwood, C. & Morris, N. (2014). Capturing personality from Facebook photos and photo-related activities: How much exposure do you need? *Computers in Human Behavior*, *37*, 162–170. doi:10.1016/j.chb.2014.04.048

Egloff, B. & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology*, *83*(6), 1441–1455. doi:10.1037/0022-3514.83.6.1441

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe Verlag.

Epskamp, S. (2019). semPlot: Path diagrams and visual analysis of various SEM packages.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... De Cock, M. (2016). Computational personality recognition in social media. *User Modelling and User-Adapted Interaction*, *26*(2-3), 109–142. doi:10.1007/s11257-016-9171-0

Fekken, G. C. & Holden, R. R. (1992). Response latency evidence for viewing personality traits as schema indicators. *Journal of Research in Personality*, *26*(2), 103–120. doi:10.1016/0092-6566(92)90047-8

Ferrando, P. J. & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525–543. doi:10.1177/0146621606295197

Ferwerda, B., Schedl, M. & Tkalcic, M. (2015). Predicting personality traits with Instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015 - EMPIRE '15* (S. 7–10). doi:10.1145/2809643.2809644

Fisher, A., Rudin, C. & Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. arXiv Preprint: 1801.01489v3

Fiske, D. W. & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, *52*(3), 217–250. doi:10.1037/h0045276

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi:10.18637/jss.v033.i01

Gerlitz, J.-Y. & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *Research Notes 4*, 1–44.

GESIS. (2018). *GESIS Panel - Standard Edition: ZA5665 Datenfile Version 26.0.0*. doi:10.4232/1.13158

Ghahramani, Z. & Jordan, M. I. (1993). Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems 6*, *6*, 120–127.

Golbeck, J., Robles, C. & Turner, K. (2011). Predicting personality with social media. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (S. 253–262). doi:10.1145/1979742.1979614

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. doi:10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment*, *4*(1), 26–42. doi:10.1037/1040-3590.4.1.26

Goretzko, D., Pham, T. T. H. & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*. doi:10.1007/s12144-019-00300-2

Greiff, S. & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. doi:10.1027/1015-5759/a000450

Guntuku, S. C., Qiu, L., Roy, S., Lin, W. & Jakhetiya, V. (2015). Do others perceive you as you want them to? Modeling personality based on selfies. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia* (S. 21–26). doi:10.1145/2813524.2813528

Hahn, E., Gottschling, J. & Spinath, F. M. (2012). Short measurements of personality - Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, *46*(3), 355–359. doi:10.1016/j.jrp.2012.03.008

Haig, B. D. (1996). Statistical methods in education and psychology: A critical perspective. *Australian Journal of Education*, *40*(2), 190–219. doi:10.1177/000494419604000206

Hartigan, J. A. & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. doi:10.2307/2346830

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

Henrich, J., Heine, S. J. & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, *466*, 29. doi:10.1038/466029a

Henrich, J., Heine, S. J. & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi:10.1017/S0140525X0999152X

Hilbig, B. E., Zettler, I., Moshagen, M. & Heydasch, T. (2013). Tracing the path from personality - via cooperativeness - to conservation. *European Journal of Personality*, *27*(4), 319–327. doi:10.1002/per.1856

Hooper, D., Coughlan, J. & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–59.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi:10.1007/bf02289447

Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Isaacowitz, D. M. (2005). The gaze of the optimist. *Personality and Social Psychology Bulletin*, *31*(3), 407–415. doi:10.1177/0146167204271599

Jain, A. K. (2008). Data clustering: 50 years beyond k-means. In W. Daelemand, B. Goethals & K. Morik (Hrsg.), *Machine Learning and Knowledge Discovery in Databases* (S. 3–4). doi:10.1007/978-3-540-87479-9_3

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: A review. *ACM computing surveys*, *31*(3), 264–323. doi:10.1145/331499.331504

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Jolliffe, I. T. (2002). *Principal component analysis* (2. Aufl.). New York: Springer.

Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. doi:10.1126/science.aaa8415

Judge, T. A., Higgins, C. A., Thoresen, C. J. & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*(3), 621–651. doi:10.1111/j.1744-6570.1999.tb00174.x

Karatzoglou, A., Smola, A. & Hornik, K. (2004). kernlab - An S4 package for kernel methods in R.

Kaspar, K. & König, P. (2011). Overt attention and context factors: The impact of repeated presentations, image type, and individual motivation. *PLoS ONE*, *6*(7), 1–15. doi:10.1371/journal.pone.0021719

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, *53*(11), 3735–3745. doi:10.1016/j.csda.2009.04.009

Kosinski, M., Stillwell, D. & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. doi:10.1073/pnas.1218772110

Kosinski, M., Wang, Y., Lakkaraju, H. & Leskovec, J. (2016). Mining Big Data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493–506. doi:10.1037/met0000105

Kotsiantis, S. B. (2007). Supervised Machine Learning: A review of classification techniques. *Informatica*, *31*(160), 249–268.

Krantz, J. H., Ballard, J. & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, and Computers*, *29*(2), 264–269. doi:10.3758/BF03204824

Kuiper, N. A. (1981). Convergent evidence for the self as prototype: The "inverted-U RT Effect" for self and other judgements. *Personality and Social Psychology Bulletin*, *7*(3), 438–443. doi:10.1177/014616728173012

Kvalseth, T. O. (1985). Cautionary note about R2. *The American Statistician*, *39*(4), 279–285. doi:10.1080/00031305.1985.10479448

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, *22*(1), 45–55. doi:10.1037/h0072400

Lenzner, T., Neuert, C. & Otto, W. (2015). Kognitives Pretesting (Version 1.1). *Mannheim, GESIS-Leibnitz-Institut für Sozialwissenschaften (SDM Survey Guidelines)*. doi:10.15465/gesis-sg_010

Lim, M. & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, *24*(3), 627–654. doi:10.1080/10618600.2014.938812

Litman, L., Robinson, J. & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. doi:10.3758/s13428-016-0727-z

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E. & Ungar, L. (2016). Analyzing personality through social media profile picture choice. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (S. 211–220).

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. doi:10.2466/pr0.1957.3.3.635

Lonsdale, C., Hodge, K. & Rose, E. A. (2006). Pixels vs. paper: Comparing online and traditional survey methods in sport psychology. *Journal of Sport and Exercise Psychology*, *28*(1), 100–108. doi:10.1123/jsep.28.1.100

Lovato, P., Bicego, M., Segalin, C., Perina, A., Sebe, N. & Cristani, M. (2014). Faved! Biometrics: Tell me which image you like and I'll tell you who you are. *IEEE Transactions on Information Forensics and Security*, *9*(3), 364–374. doi:10.1109/TIFS.2014.2298370

Mackiewicz, M. & Cieciuch, J. (2016). Pictorial personality traits questionnaire for children (PPTQ-C) - A new measure of children's personality traits. *Frontiers in Psychology*, *7*, 1–11. doi:10.3389/fpsyg.2016.00498

Magee, L. (1990). R2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, *44*(3), 250–253. doi:10.1080/00031305.1990.10475731

Marengo, D., Giannotta, F. & Settanni, M. (2017). Assessing personality using emoji: An exploratory study. *Personality and Individual Differences*, *112*, 74–78. doi:10.1016/j.paid.2017.02.037

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, *35*(2), 63–78. doi:10.1037/0022-3514.35.2.63

Martsh, C. T. & Miller, W. R. (1997). Extraversion predicts heavy drinking in college students. *Personality and Individual Differences*, *23*(1), 153–155. doi:10.1016/S0191-8869(97)00015-9

McCrae, R. . & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509–516. doi:10.1037/0003-066X.52.5.509

McCrae, R. R. & John, O. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x

McCrae, R. & Costa, P. T. (1987). Validation of the five -factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. doi:10.1037/0022-3514.52.1.81

McCrae, R. R. & Costa, P. T. (1982). Self-concept and the stability of personality: Cross-sectional comparisons of self-reports and ratings. *Journal of Personality and Social Psychology*, *43*(6), 1282–1292. doi:10.1037/0022-3514.43.6.1282

McCrae, R. R. & Costa, P. T. (2008). Empirical and theoretical status of the five-factor model of personality traits. In G. J. Boyle, G. Matthews & D. H. Saklofske (Hrsg.), *The SAGE Handbook of Personality Theory and Assessment: Volume 1 - Personality Theories and Models* (Bd. 1, S. 273–294). doi:10.4135/9781849200462.n13

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis*, *52*(1), 374–393. doi:10.1016/j.csda.2006.12.019

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037//0003-066X.50.9.741

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J. & Marlatt, G. A. (2002). Test-retest reliability of aclohol measures: Is there a difference between internet-based assessment and traditional methods? *Psychology of Addictive Behaviors*, *16*(1), 56–63. doi:10.1037//0893-164x.16.1.56

Molnar, C. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, *3*(27), 786. doi:10.21105/joss.00786

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/.

Morgan, C. D. & Murray, H. (1935). A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry*, *34*(2), 289–306. doi:10.1001/archneurpsyc.1935.02250200049005

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K. & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683–729. doi:10.1111/j.1744-6570.2007.00089.x

Nathans, L. L., Oswald, F. L. & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assesment, Research & Evaluation*, *17*(9), 1–19.

Nederhof, A. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. doi:10.1002/ejsp.2420150303

Ostendorf, F. & Angleitner, A. (2004). *Neo-Persönlichkeitsinventar nach Costa und McCrae: Neo-PI-R; Manual*. Göttingen: Hogrefe Verlag.

Oswald, F. L. & Putka, D. J. (2017). Big data methods in the social sciences. *Current Opinion in Behavioral Sciences*, *18*, 103–106. doi:10.1016/j.cobeha.2017.10.006

Ottmann, T. & Widmayer, P. (2017). *Algorithmen und Datenstrukturen* (6. Aufl.). Berlin: Springer.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, *97*(2), 307–315. doi:10.1037/0033-2909.97.2.307

Ozer, D. J. & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*(1), 401–421. doi:10.1146/annurev.psych.57.102904.190127

Pargent, F. & Gönna, J. A.-v. D. (2018). Predictive modeling with psychological panel data. *Zeitschrift für Psychologie*, *226*(4), 246–258. doi:10.1027/2151-2604/a000343

Paunonen, S. V., Ashton, M. C. & Jackson, D. N. (2001). Nonverbal assessment of the Big Five personality factors. *European Journal of Personality*, *15*(1), 3–18. doi:10.1002/per.385

Paunonen, S. V., Jackson, D. N. & Keinonen, M. (1990). The structured nonverbal assessment of personality. *Journal of Personality*, *58*(3), 481–502. doi:10.1111/j.1467-6494.1990.tb00239.x

Paunonen, S. V., Zeidner, M., Engvik, H. A., Oosterveld, P. & Maliphant, R. (2000). The nonverbal assessment of personality in five cultures. *Journal of Cross-Cultural Psychology*, *31*(2), 220–239. doi:10.1177/0022022100031002005

Peng, K.-H., Liou, L.-H., Chang, C.-S. & Lee, D.-S. (2015). Predicting personality traits of Chinese users based on Facebook wall posts. In *24th Wireless and Optical Communication Conference, WOCC* (S. 9–14). doi:10.1109/WOCC.2015.7346106

Penton-Voak, I. S., Pound, N., Little, A. C. & Perett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, *24*(5), 607–640. doi:10.1521/soco.2006.24.5.607

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*(3), 450–461. doi:10.1086/323732

Picard, R. R. & Cook, R. D. (1984). Cross-Validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583. doi:10.1080/01621459.1984.10478083

Pratama, B. Y. & Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015* (S. 170–174). doi:10.1109/ICODSE.2015.7436992

Probst, P., Bischl, B. & Boulesteix, A.-L. (2018). *Tunability: Importance of hyperparameters of Machine Learning algorithms*. arXiv Preprint: 1802.09596

Quercia, D., Kosinski, M., Stillwell, D. & Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011* (S. 180–185). doi:10.1109/PASSAT/SocialCom.2011.26

R Core Team. (2018). R: A language and environment for statistical computing. Wien, Österreich: R Foundation for Statistical Computing.

Rad, M. S., Martingano, A. J. & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. doi:10.1073/pnas.1721165115

Rammstedt, B., Kemper, C., Klein, M. C., Beierlein, C. & Kovaleva, A. (2014). Big Five Inventory 10 (BFI-10). *Zusammenstellung sozialwissenschaftlicher Items und Skalen*, 1–23. doi:10.6102/zis76

Ranger, J. & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389–406. doi:10.1177/0013164410382895

Rauthmann, J. F., Seubert, C. T., Sachse, P. & Furtner, M. R. (2012). Eyes as windows to the soul: Gazing behavior is related to personality. *Journal of Research in Personality*, *46*(2), 147–156. doi:10.1016/j.jrp.2011.12.010

Ray, S. & Page, D. (2001). Multiple instance regression. *ICML*, *1*, 425–432.

Reese, G., Proch, J. & Cohrs, J. C. (2014). Individual differences in responses to global inequality. *Analyses of Social Issues and Public Policy*, *14*(1), 217–238. doi:10.1111/asap.12032

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. Evanston, Illinois, USA: Northwestern University.

Riva, G., Teruzzi, T. & Anolli, L. (2003). The use of the internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior*, *6*(1), 73–80. doi:10.1089/109493103321167983

Roberts, B. W. & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*(1), 3–25. doi:10.1037/0033-2909.126.1.3

Rorschach, H. (1992). *Psychodiagnostik: Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experiments (Deutenlassen von Zufallsformen)* (11. Aufl.). Bern: Huber.

Rosset, S., Perlich, C. & Zadrozny, B. (2007). Ranking-based evaluation of regression models. *Knowledge and Information Systems*, *12*(3), 331–353. doi:10.1007/s10115-006-0037-3

Rouquette, A. & Falissard, B. (2011). Sample size requirements for the internal validation of psychiatric scales. *International Journal of Methods in Psychiatric Research*, *20*(4), 235–249. doi:110.1002/mpr.352

Rule, N. O., Ambady, N. & Adams, R. B. (2009). Personality in perspective: Judgmental consistency across orientations of the face. *Perception*, *38*(11), 1688–1699. doi:10.1068/p6384

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology*, *82*(1), 30–43. doi:10.1037//0021-9010.82.1.30

Saucier, G. & Goldberg, L. R. (1996). Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality*, *10*(1), 61–77. doi:10.1002/(SICI)1099-0984(199603)10:1<61::AID-PER246>3.0.CO;2-D

Schmitt, D. P., Allik, J., McCrae, R. R., Benet-Martínez, V., Alcalay, L., Ault, L., … Zupanèiè, A. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*(2), 173–212. doi:10.1177/0022022106297299

Schmitt, D. P., Realo, A., Voracek, M. & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*(1), 168–182. doi:10.1037/0022-3514.94.1.168

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), 1–16. doi:10.1371/journal.pone.0073791

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., … Lepri, B. (2017a). What your Facebook profile picture reveals about your personality. In *Proceedings of the 2017 ACM on Multimedia Conference* (S. 460–468). doi:10.1145/3123266.3123331

Segalin, C., Perina, A., Cristani, M. & Vinciarelli, A. (2017b). The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing*, *8*(2), 268–285. doi:10.1109/TAFFC.2016.2516994

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330

Shmueli, G. (2017). Analyzing behavioral big data: Methodological, practical, ethical, and moral issues. *Quality Engineering*, *29*(1), 57–74. doi:10.1080/08982112.2016.1210979

Shmueli, G. & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572. doi:10.2307/23042796

Silvia, P. J., Fayn, K., Nusbaum, E. C. & Beaty, R. E. (2015). Openness to experience and awe in response to nature and music: Personality and profound aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(4), 376–384. doi:10.1037/aca0000028

Smith, G. T. & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 300–308. doi:10.1037/1040-3590.7.3.300

Smola, A. J. & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. doi:10.1023/b:stco.0000035301.49549.88

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. doi:10.2307/1412159

Steelman, Z. R., Hammer, B. I. & Limayem, M. (2014). Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly*, *38*(2), 355–378. doi:10.25300/misq/2014/38.2.02

Steinwart, I. & Thomann, P. (2017). liquidSVM: A fast and versatile SVM package. arXiv Preprint: 1702.06899

Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. doi:10.1002/(sici)1099-0984(199909/10)13:5<389::aid-per361>3.0.co;2-a

Stone, M. (1974). Cross-validatiotory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *36*(2), 111–133. doi:10.1111/j.2517-6161.1976.tb01573.x

Tett, R. P., Jackson, D. N. & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*(4), 703–742. doi:10.1111/j.1744-6570.1991.tb00696.x

Therneau, T. M. & Atkinson, B. (2018a). rpart: Recursive partitioning and regression trees.

Therneau, T. M. & Atkinson, E. J. (2018b). An introduction to recursive partitioning using the rpart routines.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tkalcic, M., Kunaver, M., Kosir, A. & Tasic, J. (2011). Addressing the new user problem with a personality based user. In *Proceedings of the Second International Workshop on User Models for Motivational Systems: the affective and the rational routes to persuasion (UMMS)* (S. 106–111).

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology*, *3*, 1–5. doi:10.3389/fpsyg.2012.00218

Tuckey, J. W. (1950). Discussion: Symposium on statistics for the clinician. *Journal of Clinical Psychology*, *6*(1), 61–74. doi:10.1002/1097-4679(195001)6:1<1::AID-JCLP2270060102>3.0.CO;2-O

van de Ven, N., Bogaert, A., Serlie, A., Brandt, M. J. & Denissen, J. J. (2017). Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology*, *32*(6), 418–429. doi:10.1108/JMP-07-2016-0220

Van Der Heide, B., D'Angelo, J. D. & Schumaker, E. M. (2012). The effects of verbal versus photographic self-presentation on impression formation in facebook. *Journal of Communication*, *62*(1), 98–116. doi:10.1111/j.1460-2466.2011.01617.x

van Dijck, J. (2008). Digital photography: Communication, identity, memory. *Visual Communication*, *7*(1), 57–76. doi:10.1177/1470357207084865

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. doi:10.1109/72.788640

Vecchione, M., Dentale, F., Alessandri, G. & Barbaranelli, C. (2014). Fakability of implicit and explicit measures of the Big Five: Research findings from organizational settings. *International Journal of Selection and Assessment*, *22*(2), 211–218. doi:10.1111/ijsa.12070

Vigneau, E. & Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics: Simulation and Computation*, *32*(4), 1131–1150. doi:10.1081/SAC-120023882

Vinciarelli, A. & Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*, *5*(3), 273–291. doi:10.1109/taffc.2014.2330816

Volodina, A., Nagy, G. & Köller, O. (2015). Success in the first phase of the vocational career: The role of cognitive and scholastic abilities, personality factors, and vocational interests. *Journal of Vocational Behavior*, *91*, 11–22. doi:10.1016/j.jvb.2015.08.009

Walker, M. & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*(4), 609–624. doi:10.1037/pspp0000064

Wang, S., Nan, B., Rosset, S. & Zhu, J. (2011). Random Lasso. *The Annals of Applied Statistics*, *5*(1), 468–485. doi:10.1214/10-AOAS377

Wang, Z. & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, *26*(1), 98–117. doi:10.1109/MSP.2008.930649

Wei, H., Zhang, F., Yuan, N. J., Cao, C., Fu, H., Xie, X., ... Ma, W.-Y. (2017). Beyond the words: Predicting user personality from heterogenous information. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17* (S. 305–314). doi:10.1145/3018661.3018717

Wickham, H. (2011). plyr: The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag.

Wickham, H., François, R., Henry, L. & Müller, K. (2019a). dplyr: A Grammar of Data Manipulation.

Wickham, H., Hester, J. & Chang, W. (2019b). devtools: Tools to Make Developing R Packages Easier.

Wickham, H., Hester, J. & François, R. (2018). readr: Read rectangular text data.

Wickham, H. & Miller, E. (2019). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.

Wilkowski, B. M., Robinson, M. D., Gordon, R. D. & Troop-Gordon, W. (2007). Tracking the evil eye: Trait anger and selective attention within ambiguously hostile scenes. *Journal of Research in Personality*, *41*(3), 650–666. doi:10.1016/j.jrp.2006.07.003

Wintre, M. G., North, C. & Sugar, L. A. (2007). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/Psychologie Canadienne*, *42*(3), 216–225. doi:10.1037/h0086893

Wolfgang, S. (2005). An integrated model of text and picture comprehension. In R. Mayer (Hrsg.), *The Cambridge Handbook Of Multimedia Learning* (S. 72–103). doi:10.1017/CBO9781139547369.006

Woo, S. E., Chernyshenko, O. S., Longley, A., Zhang, Z. X., Chiu, C. Y. & Stark, S. E. (2014). Openness to experience: Its lower level structure, measurement, and cross-cultural equivalence. *Journal of Personality Assessment*, *96*(1), 29–45. doi:10.1080/00223891.2013.806328

Worthington, R. L. & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. doi:10.1177/0011000006288127

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi:10.18637/jss.v077.i01

Wu, Y. C. J., Chang, W. H. & Yuan, C. H. (2015). Do Facebook profile pictures reflect user's personality? *Computers in Human Behavior*, *51*, 880–889. doi:10.1016/j.chb.2014.11.014

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi:10.1177/1745691617693393

Yin, P. & Fan, X. (2001). Estimating R2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, *69*(2), 203–224. doi:10.1080/00220970109600656

Youyou, W., Kosinski, M. & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. doi:10.1073/pnas.1418680112

Yuan, M. & Lin, Y. (2006). Model selection and estimation in additive regression models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *68*(1), 49–67. doi:10.1111/j.1467-9868.2005.00532.x

Zhao, P. & Yu, B. (2004). Boosted Lasso. *UCB-STATS-678. California University Berkeley Department of Statistics*. doi:10.21236/ada473146

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi:10.1198/016214506000000735

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x