



**School of  
Engineering**

CAI Centre for  
Artificial Intelligence

## **Bachelorarbeit (Informatik)**

Was verraten Bildpräferenzen über die  
Persönlichkeit? – Eine explorative  
Datenanalyse

---

**Autoren**

---

Timos Papidas  
Yves Geistlich

---

**Hauptbetreuung**

---

Manuela Hürlimann  
Mark Cieliebak

---

**Datum**

---

07.06.2024

## **Eigenständigkeitserklärung**

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig oder gemeinsam mit den aufgeführten Gruppenmitgliedern verfasst habe.

Ich habe ausschliesslich die im Text oder Anhang angegebenen Quellen und Hilfsmittel (auch Internetseiten und generative KI-Tools) benutzt. Ich trage die Verantwortung für die Qualität des Textes sowie die Auswahl aller Inhalte und habe sichergestellt, dass Informationen und Argumente mit geeigneten wissenschaftlichen Quellen belegt bzw. gestützt werden. Schliesslich sei erwähnt, dass generative KI-Systeme bzw. KI-Tools in verschiedenen Prozessphasen dieser Arbeit zum Einsatz kamen. Im Einzelnen wurde ChatGPT für die Verbesserung der Formulierungen des Texts und das Verstehen von Inhalten verwendet.

Bei Verfehlungen aller Art treten die Paragraphen 39 und 40 (Unredlichkeit und Verfahren bei Unredlichkeit) der ZHAW-Rahmenprüfungsordnung sowie die Bestimmungen der Disziplinar massnahmen der Hochschulordnung in Kraft.

**Ort, Datum:**

**Name Studierende:**

07.06.2024, Zürich

Timos Papidas

07.06.2024, Zürich

Yves Geistlich

## **Zusammenfassung**

Die „Laufbahndiagnostik“ ist eine öffentlich zugängliche Internetseite vom Institut für angewandte Psychologie der ZHAW. Darauf finden sich diverse Fragebögen und Arbeitsmittel, die einen dabei unterstützen sollen, sich systematisch mit der beruflichen Laufbahn auseinanderzusetzen. Eine Umfrage, BFAS-G-R beschreibt das Big Five Modell, das besagt, dass eine Persönlichkeit Anhand von fünf Dimensionen beschreibbar ist. Die Ressourcenbilder sind ein Arbeitsmittel, bei dem man Bildpräferenzen angeben muss, die favorisierten Bilder kategorisiert und zum Lieblingsbild ein kurzer Text geschrieben wird. In dieser Arbeit wird versucht, herauszufinden was in den Daten der Ressourcenbilder steckt und ob sich daraus die BFAS-G-R Dimensionen vorhersagen lassen. Dazu wurden zwei Ansätze verfolgt. Der erste Ansatz versucht über gängige Clusteringmethoden eine Korrelation zwischen den beiden Datensätzen zu finden. Dies war nicht möglich, es liessen sich keine Gemeinsamkeiten zwischen den verschiedenen Datensätzen finden. Der nächste Ansatz war ein Vorhersagemodell, welches aus den Bilddaten mit Hilfe von Machine Learning Verfahren die BFAS-G-R Dimensionen vorhersagen soll. Die Ergebnisse haben hier gezeigt, dass durchaus gute Ergebnisse erzielt werden können, jedoch war die Baseline, der Mittelwert, derart gut, dass je nach Dimension kaum unterschritten werden konnte. Als Gründe werden die geringe Datenanzahl und fehlende psychologische Relevanz genannt.

## **Abstract**

„Laufbahndiagnostik“ is a publicly accessible website from the Institute of Applied Psychology at the ZHAW. It contains various questionnaires and tools designed to help you systematically examine your professional career. One survey, BFAS-G-R, describes the Big Five model, which states that a personality can be described on the basis of five dimensions. The resource pictures are a working tool in which you have to indicate picture preferences, categorize your favourite pictures and write a short text about your favourite picture. This paper attempts to find out what is contained in the data of the resource pictures and whether the BFAS-G-R dimensions can be predicted from them. Two approaches were pursued for this purpose. The first approach tried to find a correlation between the two data sets using common clustering methods. This was not possible, as no similarities could be found between the different data sets. The next approach was a prediction model that uses machine learning methods to predict the BFAS-G-R dimensions from the image data. The results here showed that good results can be achieved, but the baseline, the mean value, was so good that it could hardly be undercut depending on the dimension. The reasons given were the small number of data and the lack of psychological relevance.

# 1. Contents

1.	Contents .....	3
2.	Einleitung .....	5
2.1	Laufbahndiagnostik .....	5
2.2	Werkzeuge .....	6
2.2.1	Persönliche Daten .....	6
2.2.2	Gemütslage .....	6
2.2.3	Ressourcenbilder .....	7
2.2.4	BFAS-G und BFAS-G-R .....	8
2.3	Fragestellung .....	10
3.	Related work .....	10
3.1	Unveröffentlichtes Paper vom CAI .....	10
3.2	Ähnliche Arbeiten .....	12
4.	Daten .....	14
4.1	Rohdaten .....	14
4.2	Preprocessing .....	15
4.2.1	Skalierung der BFAS-Datensätze .....	15
4.2.2	Umpolung revidierter Features .....	16
4.2.3	Versionstrennung des Ressourcenbilderdatensatzes .....	17
4.2.4	Säuberung und Vereinigung .....	19
4.3	Datenanalyse .....	20
4.3.1	Stichprobe .....	20
4.3.2	BFAS .....	21
4.3.3	Ressourcenbilder .....	25
4.3.4	Zusammenfassung .....	27
5.	Clustering .....	28
5.1	Methoden .....	28
5.1.1	Clustering Algorithmen .....	28
5.1.2	Dimensionality Reduction .....	28
5.1.3	Clusteroptimierung .....	29
5.1.4	Cluster Correspondence Analysis .....	30
5.2	Ergebnisse .....	30
6.	Prediction Models .....	34
6.1	Vorgehensweise .....	34

6.2	Ergebnisse .....	35
6.3	Weitere Analysen.....	39
7.	Diskussion und Ausblick .....	41
8.	Literaturverzeichnis.....	43
9.	Abbildungsverzeichnis .....	44
10.	Technische Dokumentation .....	46
11.	Anhang .....	46

## 2. Einleitung

### 2.1 Laufbahndiagnostik

Die Grundlage dieser Arbeit legte das Institut für angewandte Psychologie (IAP) der ZHAW mit der eigens entwickelten Laufbahndiagnostik-Plattform. Diese Plattform ist über die Website [www.laufbahndiagnostik.ch](http://www.laufbahndiagnostik.ch) öffentlich verfügbar. Sie beinhaltet diverse Werkzeuge zur Laufbahnberatung, darunter Fragebogen zu den folgenden Themen: Berufliche Laufbahn, Persönlichkeit, Motive, Interesse und Wohlbefinden. Diese dienen nicht nur der Datensammlung, sondern auch der Quantifizierung psychologischer Faktoren. Zum Beispiel wird Laufbahn Adaptabilität – also Faktoren, welche eine positive berufliche Laufbahnentwicklung begünstigen – mit der Career Adapt-Abilities Skala (CAAS) gemessen [1]. Bei diesen Fragebögen handelt es sich um psychometrische Instrumente, die auf Forschungserkenntnissen basieren.

Darüber hinaus bietet die Plattform diverse Arbeitsmittel. Im Gegensatz zu den Fragebogen geht es bei den Arbeitsmitteln weniger um die Quantifizierung der Psyche, sie dienen eher als Anregung zur Selbstreflexion [2]. Beispielsweise können Klienten hier berufliche Werte auswählen, welche ihnen besonders wichtig sind. Weitere Funktionen der Plattform sind Identitätskarten, mit denen gewonnene Erkenntnisse festgehalten werden können und das Cockpit, wo die ausgefüllten Fragebögen und Arbeitsmittel inklusive eines darauf basierenden psychologischen Profils eingesehen werden können. Ausserdem kann hier auch das Nutzerprofil angepasst werden, wo persönliche Daten, welche für die Karriere relevant sind, wie zum Beispiel der höchste Bildungsabschluss, oder die aktuelle Berufsposition, erfasst werden.

Die Laufbahndiagnostik ist aber nicht nur eine Sammlung von diversen psychologischen Werkzeugen. Die Idee dahinter ist, dass das psychologische Profil, welches mithilfe der Fragebogen und Arbeitsmittel angefertigt wird, mit einer persönlichen Beratung durch Experten verbunden werden kann. Per E-Mail kann das Profil an eine Beratungsperson weitergeleitet werden, um dazu kostenlos Feedback zu bekommen. So verbindet die Laufbahndiagnostik psychologische und Karrieredaten mit Selbstreflexion und dem Wissen einer Beratungsperson, um die Klienten bei der Planung und Gestaltung ihrer beruflichen Laufbahn zu unterstützen [3].

Diese Arbeit basiert konkret auf zwei psychologischen Werkzeugen der Laufbahnberatung: Dem Fragebogen „BFAS-G“ beziehungsweise „BFAS-G-R“, und dem Arbeitsmittel „Ressourcenbilder“. Letzteres wird im weiteren Verlauf dieser Arbeit ebenfalls als Fragebogen anstatt als Arbeitsmittel bezeichnet.

## 2.2 Werkzeuge

### 2.2.1 Persönliche Daten

Um die Werkzeuge der Laufbahndiagnostik zu nutzen, müssen Klienten zuerst persönliche Daten erfassen. Darunter sind zwingend erforderliche Angaben, wie Name und Geburtstag, und freiwillige Angaben, wie das Arbeitsland. Die persönlichen Daten sollen die aktuelle berufliche Situation der Klienten erfassen. Sie dienen als Kontext, welcher nebst den Ergebnissen der weiteren Werkzeuge für die Beratungspersonen nützlich sein kann.

In dieser Arbeit wurden folgende persönliche Daten verwendet:

- Geschlecht
- Alter
- Berufsposition, beispielsweise «Sachbearbeiter/-in»
- Berufsstatus, beispielsweise «Teilzeitangestellt»
- Berufssektor, beispielsweise «Gesundheits- und Sozialwesen»
- Land
- Arbeitskanton
- Bildungsstand

### 2.2.2 Gemütslage

Da die Gemütslage der Befragten deren Antwortverhalten beeinflussen kann, beginnen alle Fragebogen mit einer Einschätzung der aktuellen Gemütslage durch die Lebender Emoticon PANA Matrix (LE-PANA-M). Die Matrix vereint positive Aktivierung (PA), negative Aktivierung (NA) und Valenz zu einem 2x2 Diagramm [4].



Abbildung 1: Screenshot LE-PANA-M, Erfassen der Gemütslage

Nach der Einschätzung mittels LE-PANA-M sollen Befragte ausserdem ihre Gemütslage kurz in eigenen Worten beschreiben. Am Ende jedes Fragebogens geben Befragte an, ob sie ihn ernsthaft oder nicht ernsthaft ausgefüllt haben.

### 2.2.3 Ressourcenbilder

Die Ressourcenbilder dienen als ein Tool im Bereich der Beratung und des Coachings, das darauf abzielt, Personen bei der Identifizierung relevanter Themen, Bedürfnisse und Motive für ihre berufliche und persönliche Entwicklung zu helfen. Durch die Auswahl eines persönlichen Ressourcenbildes und das Verfassen einer dazugehörigen Geschichte kann man sowohl explizite als auch implizite Aspekte ihrer Persönlichkeit und Identität erforschen. Aus theoretischer Sicht integriert dieses Tool Konzepte der Career Construction Theorie und der Persönlichkeits-System-Interaktionen (PSI-Theorie), wobei es besonderen Wert auf die intuitive und emotionale Verarbeitung legt. Es bietet einen strukturierten Ansatz, um über visuelle und narrative Elemente tiefere Einsichten in das eigene Selbst und die berufliche Laufbahn zu gewinnen. Nach ersten Versuchen wurden mehr Bilder mit sozialen Situationen einbezogen, damit das Zwischenmenschliche mehr zum Ausdruck kommt.

#### Ressourcenbilder

3 von 80



👎 Spricht mich nicht an

Neutral

👍 Spricht mich an

Abbildung 2: Beispiel für ein Ressourcenbild

Der Ressourcenbilderfragebogen besteht je nach Version aus 70 oder 80 Bildern. Die Befragten sollen jedes Bild positiv, neutral oder negativ bewerten.

Daraufhin müssen die Befragten all die von ihnen positiv gewerteten Bilder nach eigenen gewählten Gruppen gruppieren. Schliesslich sollte das bevorzugte Bild ausgewählt und gemäß den folgenden Vorgaben ein kurzer Text dazu verfasst werden:

*„Bitte versuchen Sie jetzt, eine Geschichte darüber zu erfinden, was in diesem Bild gerade vorgeht und was vorher geschehen ist. Was denken und fühlen die Leute auf dem Bild und welche Absichten verfolgen sie.4 Versuchen Sie eine abgeschlossene Geschichte zu erzählen, die einen Anfang, ein Mittelstück, und ein Ende hat. Es gibt keine ‚richtigen‘ oder ‚falschen‘ Geschichten. Schreiben Sie also die Geschichte darüber, was Ihnen als erstes in den Sinn kommt. Rechtschreibung und Grammatik sind völlig unwichtig. Es ist dagegen wichtig, dass Sie sich vorzustellen versuchen, was in diesem Bild vorgeht, und darüber eine Geschichte schreiben.“*

*Die folgenden Fragen sollen Ihnen beim Schreiben der Geschichte helfen:*

- Was passiert gerade?



- Wer sind die abgebildeten Leute?
- Was passierte vorher?
- Was denken und fühlen die Leute?
- Welche Absichten und Wünsche verfolgen sie?
- Was wird als nächstes passieren?

Betrachten Sie das Bild während 15–20 Sekunden und schreiben Sie dann die Geschichte. Das Verfassen der Geschichte sollte nicht länger als 10 Minuten dauern. Ihre Geschichte:” [5]BFAS-G und BFAS-G-R Ihre Geschichte:” [5]

#### 2.2.4 BFAS-G und BFAS-G-R

Die theoretische Grundlage dieser Fragebogen ist das Big Five Modell. Dieses Modell versucht, die Persönlichkeit durch fünf Dimensionen Extraversion, Neurotizismus, Offenheit/Intellekt, Gewissenhaftigkeit und Verträglichkeit zu beschreiben [6].

Neue Forschungsergebnisse suggerierten allerdings pro Persönlichkeitsdimension 2 biologisch unterschiedliche Aspekte. Deshalb wurde in den 2000er Jahren das Big Five Modell um zwei Persönlichkeitsaspekte pro Dimension erweitert. Dieses neue Modell, genannt „Big Five Aspect“ Scales (BFAS) ermöglicht eine detailliertere Erfassung der Persönlichkeit in 10 Aspekten mittels 100 Fragen [7]. Im Kontext des Fragebogens als psychometrisches Werkzeug

Dimension	Aspekte	Beschreibung
Extraversion	Durchsetzungsvermögen	Personen mit hohen Werten in Extraversion sind abenteuerlustig, aktiv, gesprächig, gesellig, herzlich und optimistisch.
	Enthusiasmus	
Neurotizismus	Volatilität	Personen mit hohen Werten neigen dazu, nervös, ängstlich und traurig zu sein. Sie sind weniger in der Lage, ihre Impulse zu kontrollieren und in Stresssituationen angemessen zu reagieren
	Rückzug	
Offenheit/Intellekt	Offenheit für Erfahrungen	Personen mit hohen Werten zeichnen sich durch eine hohe Wertschätzung für neue Erfahrungen aus, sind wissbegierig und haben vielfältige kulturelle Interessen.
	Intellekt	
Gewissenhaftigkeit	Fleiss	Personen mit hohen Werten sind zielstrebig, ausdauernd, ordnungsbewusst und zuverlässig.
	Ordnung	
Verträglichkeit	Mitgefühl	Personen mit hohen Werten sind altruistisch, verständnisvoll, wohlwollend, kooperativ und streben nach Harmonie.
	Höflichkeit	

Tabelle 1: Dimensionen und Aspekte von BFAS-G

werden diese Fragen auch Items genannt.

BFAS-G ist die Übersetzung der BFAS in die deutsche Sprache [8]. Bei BFAS-G-R handelt es sich um eine Weiterentwicklung des Fragebogens BFAS-G durch das Institut für Angewandte Psychologie (IAP) der ZHAW. Die wichtigsten Unterschiede zwischen BFAS-G und BFAS-G-R sind folgende:

- Alle Fragen in BFAS-G sind so gepolt, dass ein hoher Wert im Fragebogen einer hohen Ausprägung der jeweiligen Eigenschaft entspricht. Bei BFAS-G-R ist das nicht der Fall, bei gewissen Items wurde also die Polung umgekehrt.
- In BFAS-G-R wurde wo immer möglich auf Negationen verzichtet. Auch das bedeutet einige Umpolungen.
- Die Fragen in BFAS-G-R werden auf einer diskreten 5-Stufigen Likert-Skala beantwortet, während BFAS-G eine diskrete 7-Stufige Likert-Skala gebraucht.

Diese Änderungen basieren auf praktischen Erfahrungen in der Laufbahnberatung, wo die Negationen und inkonsistente Polung der Fragen teilweise zu Verwirrung bei den Klienten führten. Die 7-Stufige Likert-Skala erwies sich als ungeeignet, da die Stufen zwischen Mitte und Extrema sich nur schwer einschätzen und benennen liessen [9].

Bitte geben Sie an, inwiefern die Aussagen auf Sie zutreffen:*	trifft nicht zu (1)	trifft eher nicht zu (2)	teils/teils (3)	trifft eher zu (4)	trifft zu (5)
Ich übernehme Verantwortung.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich werde schnell ärgerlich.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich kann die Schönheit der Natur sehr geniessen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich führe meine Vorhaben zu Ende.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich interessiere mich sehr für die Probleme anderer.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Es fällt mir leicht, neue Freundschaften zu schliessen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich fühle mich häufig deprimiert.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich habe eine gute Auffassungsgabe.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich halte Ordnung mit meinen Sachen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Autoritäten begegne ich stets mit viel Respekt.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Abbildung 3: BFAS-G-R Frageschema

Ausserdem bezieht sich BFAS-G-R stark auf die Cybernetic Big Five Theory (CB5T) von DeYoung [10], „welche Persönlichkeit aus einer systemtheoretischen Perspektive betrachtet und Eigenschaften als Ausdruck des kybernetischen Mechanismus der Zielerreichung auffasst, nachdem Menschen grundsätzlich funktionieren [9].“

Auf der Seite der Laufbahnberatung ist nur der BFAS-G-R Fragebogen zugänglich, da er den BFAS-G ersetzt hat. Für diese Arbeit wurden allerdings sowohl Daten des veralteten BFAS-G als auch des neuen BFAS-G-R verwendet.

## 2.3 Fragestellung

Diese Arbeit wurde als Explorative Datenanalyse ausgeschrieben, was eine offene Arbeit impliziert. Im nachfolgenden Kapitel wird auf eine bereits durchgeführte Arbeit eingegangen, welche bereits den Textaspekt der Ressourcenbilder untersucht hat. In dieser Arbeit hat die Textanalyse nicht wie gewünscht funktioniert, weshalb man sich von Anfang an nur auf die Bilddaten beschränkt hat.

Zu Beginn der Arbeit war noch nicht klar, was genau umgesetzt werden sollte, wie genau unsere Fragestellung lautet. Auf Grund dessen, wurde zuerst etwas Zeit investiert, um die Daten zu verstehen. Daraus formulierte sich dann folgende Fragestellung.

- Was sind die Fragen des BFAS-Fragebogen, wie sehen die Ressourcenbilder aus?
- Welche Leute haben diesen Fragebogen ausgefüllt?
- Gibt es beispielsweise Unterschiede im Geschlecht, was die Resultate betrifft?
- Sind die verschiedenen BFAS- und Ressourcenbilderversionen kompatibel?

Daraus wurde folgende Fragestellung abgeleitet:

Was steckt in den Daten der Ressourcenbilder drin und ist es möglich, aus den Daten der Ressourcenbilder die BFAS Dimensionen hervorzusagen?

Um diese Frage zu beantworten, wurden mehrere Ansätze verfolgt, die in den späteren Kapiteln erläutert werden.

## 3. Related work

### 3.1 Unveröffentlichtes Paper vom CAI

Dieser Arbeit geht bereits ein Projekt der ZHAW voraus. Darin wurde untersucht, ob“ KI idiografische Erzählungen mit nomothetischen Fragebogendaten verbinden kann“. In diesem Projekt wurde die gleiche Datengrundlage verwendet, jedoch ging es bei diesem Projekt darum, dass Persönlichkeit aus den formulierten Kurztexten hervorgesagt wird, während wir dazu die Bildpräferenzen benutzten.

In der Arbeit wurden zwei unterschiedliche Ansätze verfolgt. Im ersten wurde ein „traditionelles“ Modell verwendet, welches Features, wie die TF-IDF (term frequency – inverse document frequency) und LDA (latent dirichlet allocation), aus den text-basierten Daten extrahiert. danach mit einem Supervised prediction model trainiert wurde. Der zweite Ansatz benutzte neuste Technologie an Large Language Models. Konkret wurde das kommerziell verfügbare GPT 3.5 Model von OpenAI verwendet. Über API-Call wurden dann die ganzen Texts an das Modell gegeben und es wurde verlangt, dass es eine Einschätzung trifft. Zuvor wurde dem Modell erklärt, was die unterschiedlichen Kategorien der Persönlichkeit bedeuten.

Die Ergebnisse der Motivvorhersage mit traditionellen Ansätzen zeigten, dass die Regressionsergebnisse mit den verschiedenen Konfigurationen nur minimale Unterschiede gegenüber der Baseline vorweisen. Keiner der verwendeten Ansätze übertraf dabei die gewählte Baseline, den Median. Die Klassifikationsergebnisse schnitten besser ab als eine zufällige Baseline, jedoch waren sie schlechter als die adaptive Baseline und die Mehrheitsklasse.

Die Ergebnisse mit der Chat GPT Vervollständigung lieferte ähnliche Ergebnisse, hier lieferten sowohl die Regressionen und die Klassifikationen ebenfalls Ergebnisse, die keine Verbesserungen zu den Baselines zeigen.



Abbildung 4: Ergebnisse der klassischen Textanalyseverfahren mit Supervised Prediction Models auf BFAS Dimensionen

Es werden folgende Gründe für die schlechten Ergebnisse erläutert:

- Anforderungen an AI-Algorithmen: Möglicherweise war die Datenmenge zu gering oder die Texte zu kurz, um zuverlässige Vorhersagen zu treffen.
- Kontextabhängigkeit der Erzählungen: Die Geschichten könnten stark von den Bildern beeinflusst sein, die zur Erstellung der Geschichten verwendet wurden, und weniger von den tatsächlichen Motiven oder Persönlichkeitsmerkmalen der Personen.
- Implizite und explizite Facetten: Geschichten erfassen sowohl implizite als auch explizite Facetten der Persönlichkeit, während Fragebögen nur explizite Facetten abdecken.
- Idiographische Natur der Persönlichkeit: Die Beziehung zwischen verschiedenen Ebenen des Selbst könnte selbst idiographisch sein und daher schwer in nomothetische Modelle zu überführen sein.

Auf Grund dessen wurde zu Beginn unserer Arbeit entschieden, die Texte nicht zu nutzen und sich nur auf die Bildpräferenzen zu beschränken. [11]

### 3.2 Ähnliche Arbeiten

Bei der Recherche wurde versucht, noch einen Schritt zurückzugehen und herauszufinden, was es im Bereich der Persönlichkeit schon für Arbeiten gab, die insbesondere in das Feld der Vorhersagen oder der Korrelation zwischen den Dimensionen und maschinell erfassten Informationen, wie in unserem Fall Bilderpräferenzen oder selbst geschriebene Texte. Dieses Feld nennt sich „Personality Computing“ und ist ein stetig wachsender Bereich. Wie in unserem Fall sind dabei der Grossteil der Arbeiten mit sogenannten „self-assessed“, also eine Selbsteinschätzung der eigenen Persönlichkeit. Es gibt die Methode, wo man sich von anderen Personen bewerten lässt, dies wird aber in diesem Kapitel nicht berücksichtigt. [12] [13]

Ebenfalls diskutiert wird die Rolle von Big Data im Zusammenhang mit der Persönlichkeitserforschung. Es wird argumentiert, dass traditionelle idiographische Ansätze, die sich auf die Einzigartigkeit des Individuums konzentrieren, durch Big Data wiederbelebt werden können. Dabei werden methodische, ethische und rechtliche Herausforderungen betont, die durch den Einsatz großer Datenmengen entstehen. Das Paper stellt verschiedene methodische Ansätze vor, darunter die Gruppe Iterative Multiple Model Estimation (GIMME) und betont die Bedeutung individueller Verhaltensmuster. Abschließend wird betont, dass Big Data das Potenzial hat, personalisierte Interventionen in der Persönlichkeitsforschung zu ermöglichen und somit zu einem tieferen Verständnis individueller Unterschiede beizutragen. [12]

Der wohl bekannteste, vergleichbare Test aus Bildbewertungen Informationen über die Persönlichkeit zu gewinnen, ist der wahrscheinlich der Rorschachtest von Hermann Rorschach aus dem Jahre 1920. Dabei wird dem Patient nacheinander zehn verschiedene Bilder gezeigt, auf denen Tintenflecken zu sehen sind. Danach wird das Subjekt gefragt, was auf dem Bild zu sehen ist. Der Psychologe kann aufgrund der Antwort, Antwortzeit und Handhabung der Bilder die folgenden fünf Aspekte auswerten:

- Lokalisation (Ganzheitliche oder detailorientierte Deutung?)
- Determinanten (Form, Farbe, Schatten)
- Häufigkeit der Antworten (im Vergleich zu anderen Testpersonen: Banal oder originell?)
- Inhalt
- Besondere Phänomene (zum Beispiel Starrheit (Stupor), verlängerte Reaktionszeit)

Daraus lassen sich vier Persönlichkeitstypen (introversiv, extroversiv, koartiert und ambiäqual) ausmachen. Dieser Test ist nicht von einem Computer gemacht, sondern einem Psychologen, der die Antworten einschätzt. [14]

Die Unterschiede zu unserer Arbeit sind klar: Zum einen verwenden die Ressourcenbilder ganz andere Bilder, zum anderen werden andere Persönlichkeitstypen vorausgesagt.

Ein grosses Feld, in dem bereits viel erforscht ist, sind die Zusammenhänge der Big Five Dimensionen und Social Media. Eine Studie untersuchte dabei 22398 Bilder von Instagram von 113 Probanden. Das Ziel dabei war, anhand der Filter, die auf die Instagram Bilder angewendet wurden, die Persönlichkeit hervorzusagen. Es wurde mit einem Regressions Model, einer 10-fold Crossvalidation und verschiedenen Algorithmen gearbeitet. Zur Beurteilung wurde root-mean-square-error (RSME) verwendet. Als Dabei kamen Ergebnisse heraus, die eine Korrelation zeigen. So verwenden zum Beispiel Leute, die einen hohen Score im Neurotizismus erzielen Filter, die die Helligkeit erhöhen. [15]

Eine Studie beispielsweise versuchte aus einzelnen Portraits auf die fünf Dimensionen zu schliessen und erreichte dabei eine durchschnittliche Genauigkeit von 65.86%. Das Modell nimmt nur ein einzelnes Bild, extrahiert davon die Merkmale und spielt sie in das Model ein, welches am Schluss eine Kategorisierung vornimmt. Das Modell klassifiziert pro Dimension eine binäre Klasse, welches das Modell vorhersagt. Der Grund dafür ist, dass als Baseline ein Menschliches Urteilsvermögen gewählt wurde und es für Menschen einfacher ist, etwas binär zu kategorisieren, verglichen mit einem Zahlenwert. Ausserdem verringert sich so die Komplexität und die Ergebnisse werden interpretierbarer. [6]

Eine weitere Studie zeigte den Zusammenhang zwischen den Big-Five Dimensionen und Smartphone Nutzungsdaten. Dabei wurden signifikante Korrelation gefunden. Beispielsweise waren Extraversion und Neurotizismus gut vorhersagbar durch Kommunikations- und Sozialverhalten, während Gewissenhaftigkeit mit Indikatoren der täglichen Routine zusammenhing. [16]

Eine Arbeit, die relativ nah an unser Gebiet kommt, beschreibt den Ansatz aus der Rekrutierung von neuen Mitarbeitern. Es wird ein bildbasierter Test verwendet, bei dem Kandidaten 5-10 Bilder beschreiben sollten. Die Ergebnisse wurden dann gebraucht, um die Persönlichkeitseigenschaften zu bestimmen. Es liegen in dieser Studie keine Ergebnisse vor, durch die beschriebenen theoretischen Grundlagen und methodischen Schritte wird allerdings nahegelegt, dass der Ansatz funktioniert. [17]

Es gibt viele Versuche, die zeigen, dass es funktionieren kann. Dabei gibt es Versuche, die mit einer Klassifizierung arbeiten und solche die mit einer Regression arbeiten, wobei die Klassifizierung häufiger auftritt. Häufige Algorithmen, die vorkommen sind beispielsweise CNN, Lasso/Lineare Regressionen und Random Forests.

## 4. Daten

### 4.1 Rohdaten

Die Rohdaten wurden vom IAP als CSV-Dateien bereitgestellt.

Dateiname	Inhalt	Zeilen	Spalten
bfas.csv	BFAS-G Daten, persönliche Daten	911	156
bfas_g_r.csv	BFAS-G-R Daten, persönliche Daten	809	156
2024-02-11_resource_diagnostic_data_anonymized.csv	Ressourcenbilderdaten, persönliche Daten	2117	236

Tabelle 1: Darstellung der zur Verfügung gestellten Daten

Die beiden BFAS-Datensätze beinhalten jeweils 100 Spalten mit numerischen Werten, welche den Items des Fragebogens entsprechen. Diese Spalten werden im weiteren Verlauf der Arbeit auch als Features bezeichnet, um sie von den Aspekten und Dimensionen zu distanzieren. Die Features sind jeweils nach dem zugehörigen Item im Fragebogen benannt. Der Name besteht aus je einem Zeichen für die zugehörige Dimension bzw. den zugehörigen Aspekt, einem Unterstrich und einer zweistelligen Itemnummer, zum Beispiel: AP\_04 steht für Dimension «Agreeability» und Aspekt «Politeness». Auf Deutsch übersetzt: «Verträglichkeit» und «Höflichkeit»

Weiter sind in 15 Spalten sind die Werte für die fünf Dimensionen und 10 Aspekte hinterlegt, deren Namensgebung ähnlich funktioniert, wobei die Itemnummer und bei den Spalten der Dimensionen zusätzlich das Zeichen des Aspekts wegfallen, zum Beispiel: O für die Dimension «Offenheit/Intellekt», OI für den Aspekt «Intellekt». Hinzu kommt eine Spalte für die user\_id, die einzigartige Identität jeder befragten Person. Die user\_id ist mit dem Profil der Laufbahndiagnostikplattform gekoppelt. Auf den restlichen 40 Spalten befinden sich persönliche Informationen und administrative Informationen, beispielsweise darüber, wann der User erstellt wurde. Die Werte der BFAS-Items sind in ganzen Zahlen erfasst mit Wertebereichen von 1-7 im BFAS-G Datensatz, respektive 1-5 im BFAS-G-R Datensatz. Die Big-Five Dimensionen und Aspekte sind in Gleitkommazahlen auf demselben Wertebereich des Jeweiligen Datensatzes erfasst.

Der Ressourcenbilderdatensatz beinhaltet beide Versionen des Fragebogens. Die Ressourcenbilderdaten von Version 1 sind auf 70 Spalten eingetragen, die der zweiten Version auf 80 separaten Spalten, was insgesamt 150 Spalten numerischer Ressourcenbilderdaten ergibt. Die Spalten der jeweiligen Version folgen unterschiedlichen Namensschemas aber beziehen sich auf den Namen des zugehörigen Ressourcenbildes, zum Beispiel: «landschaft-mv-03\_14\_Mitternachtssonne\_Ausblick.jpg» in Version 1 ist äquivalent zu «la08\_mitternachtssonne\_ausblick\_mv.jpg» in Version 2. Das erste Wort bzw. Kürzel des Namens steht für die Kategorie des Ressourcenbildes. Für Version 2 wurden die gleichen Bilder wie für Version 1 verwendet, aber zusätzlich wurden noch 12 neue Bilder hinzugefügt. Es wurden mehr Bilder mit sozialen Situationen einbezogen, damit das zwischenmenschliche mehr zum Ausdruck kommt. Die restlichen Spalten beinhalten user\_id, persönliche und administrative Daten, und weitere Daten zu den Ressourcenbildern wie z.B. die Kurzgeschichte zu ihrem Lieblingsbild. Die 150 Spalten Ressourcenbilderdaten sind jeweils ganze Zahlen mit den Werten -1, 0, oder 1.

## 4.2 Preprocessing

Die CSV-Dateien wurden mit Python eingelesen und zu Pandas-Dataframes konvertiert. Anschliessend wurden die Dataframes folgendermassen bereinigt und zur Analyse und zum Trainieren von Modellen vorbereitet:

- Nach Absprache mit dem Experten wurde befunden, dass die beiden BFAS Datensätze miteinander kompatibel sind. Die BFAS-Datensätze wurden einander angeglichen und anschliessend vereint, womit die Datenmenge erhöht wurde.
- Da die beiden Versionen des Ressourcenbilderfragebogens nicht die gleiche Anzahl Bilder haben, sind sie nicht miteinander kompatibel. Deshalb wurde entschieden, dass der Datensatz so aufgeteilt werden muss, dass jede Version in einem eigenen Datensatz vorhanden ist.
- Die Datensätze wurden gesäubert, indem unbrauchbare oder unerwünschte Spalten entfernt wurden.
- Der vereinte BFAS-Datensatz wurde mit dem Bilderdatensatz vereint.

Die Teilprozesse des Preprocessing werden nun einzeln beleuchtet.

### 4.2.1 Skalierung der BFAS-Datensätze

Es wurde entschieden, dass BFAS-G an BFAS-G-R angeglichen werden sollte, da bei BFAS-G-R alle Items so gepolt sind, dass hohe Werte einer hohen Ausprägung des Aspektes / der Dimension darstellen. Die Berechnung der Aspekte und Dimensionen ist somit trivial, da man nur den Durchschnitt der zugehörigen Features berechnen muss.

Zuerst wurde bei beiden Datensätzen sichergestellt, dass eine user\_id jeweils nur einmal im Datensatz vertreten war. Dies vermeidet Mehrdeutigkeiten und

Die BFAS-G und BFAS-G-R Daten sind verschieden skaliert: BFAS-G auf einer ganzzahligen Skala von 1 bis 7, BFAS-G-R auf einer ganzzahligen Skala von 1 bis 5. Das bedeutet, dass die Skalen einander angeglichen werden mussten, um die beiden Versionen des Fragebogens zu vereinen. Dazu wurde auf alle Werte in BFAS-G folgende Funktion angewendet:

$$F(x) = \frac{4(x - 1)}{6} + 1$$

Der Wertebereich wurde also folgendermassen angepasst:

BFAS-G	BFAS vereint	BFAS-G-R
1	1	1
2	1.666	
	2	2
3	2.333	
4	3	3
5	3.666	
	4	4
6	4.333	
7	5	5

Tabelle 2: Mapping der Skalierungen von BFAS-G und BFAS-G-R auf den vereinten Fragebogen



Alternativ hätte man beispielsweise die Werte 2 und 3 aus BFAS-G auch kombinieren können, aber eine solche Operation verändert die Daten stark, weswegen dagegen entschieden wurde. Mit der gewählten Methode werden die relativen Abstände der vorherigen Skala beibehalten. Allerdings hat sie zur Folge, dass die BFAS-Features nicht mehr alle ganzzahlig sind, sichtbar in Tabelle 2: Mapping der Skalierungen von BFAS-G und BFAS-G-R auf den vereinten Fragebogen. Unter Umständen könnte das ein Nachteil sein, aber für diese Arbeit sind primär nicht einzelne Features, sondern die Big Five Aspekte und Dimensionen interessant. Da sich aus den Werten mehrerer Features zusammensetzen, sind sie sowieso nicht ganzzahlig, somit fällt dieser Nachteil nicht ins Gewicht.

#### 4.2.2 Umpolung revidierter Features

Für BFAS-G-R wurden 63 von 100 Items in BFAS-G angepasst. Es gab zwei Arten von Anpassungen:

- Umpolung des Items, sodass beispielsweise ein Wert von 7, ein Maximalwert in BFAS-G, einem Minimalwert, also 1, in BFAS-G-R entspricht. Dies betraf 46 der 63 geänderten Items.
- Umformulierung des Items ohne Umpolung.

Diese Anpassungen wurden vom IAP in der Excel Tabelle «BFAS-G-R\_Items zwei Versionen.xlsx» dokumentiert und uns bereitgestellt, sodass klar war, welche Items wie angepasst wurden.

Item Text alt (BFAS-G)	Polung r = reversed	Item_ Label_neu BFAS-G-R	Item Text neu BFAS-G-R	Change Y/N
Ich respektiere Autoritäten.		AP_01_R	Autoritäten begegne ich stets mit viel Respekt.	Y
Ich fühle mich anderen überlegen.	r	AP_02_R	Ich fühle mich anderen nicht überlegen.	Y
Es ist mir unangenehm, aufdringlich zu erscheinen.		AP_03	Es ist mir unangenehm, aufdringlich zu erscheinen.	N

Abbildung 5: Ausschnitt aus «BFAS-G-R\_Items zwei Versionen.xlsx». Zu sehen sind drei Items, von denen zwei angepasst wurden, siehe Spalte «Change». Bei einem Item wurde ausserdem die Polung umgekehrt, siehe Spalte «Polung r = reversed».

Auf allen Items aus BFAS-G, welche in der Dokumentation als «Polung reversed» markiert sind, muss also folgende Funktion angewendet werden, wenn man sie an ihr Gegenstück in BFAS-G-R angleichen will:

$$F(x) = N + 1 - x$$

Wobei N für das Maximum des Wertebereichs des Datensatzes steht. Bei BFAS-G ist N also 7, bei BFAS-G-R hingegen 5.

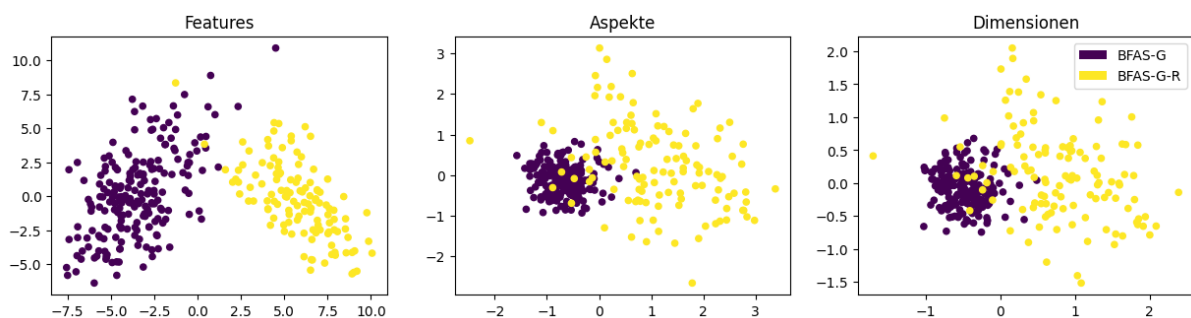


Abbildung 6: Sichtbare Unterschiede zwischen BFAS-G und BFAS-G-R aufgrund fehlerhafter Umpolung

Dieser Umpolungsschritt ist eine potenzielle Fehlerquelle, denn die Daten sollten nur einmal umgepolt werden, da eine zweifache Umpolung in den ursprünglichen Werten resultiert. Bei

dieser Arbeit wurde genau dieser Fehler gemacht, weil irrtümlich unter der Annahme gearbeitet wurde, dass die Rohdaten in der originalen Polung des jeweiligen Fragebogens übermittelt wurden. Tatsächlich war dies aber nicht der Fall, die erhaltenen BFAS-G Daten waren bereits im Vorherein umgepolt worden. Durch die erneute Umpolung während des Preprocessing wurde dies wieder rückgängig gemacht. Dieser Fehler hatte zur Folge, dass die zwei Datensätze sich stark voneinander unterschieden.

Auf Anhieb ist nicht direkt sichtbar, ob ein Datensatz umgepolt wurde oder nicht. Wenn potenziell eine Umpolung gemacht werden muss, empfiehlt sich deshalb:

1. Sich bei der Quelle der Daten zu vergewissern, ob bereits eine Umpolung stattgefunden hat.
2. Die Durchschnittswerte der Features der Datensätze vor und nach der Umpolung zu vergleichen. Verschieden gepolte Datensätze unterscheiden sich stärker als gleich gepolte.

Nach einer Rücksprache mit dem Experten wurde klar, dass die BFAS-G Rohdaten bereits in umgepolter Form übermittelt worden waren, weshalb dieser Schritt im Preprocessing schlussendlich weggelassen wurde. Nun da sichergestellt war, dass die Polung aller Features in beiden Datensätzen übereinstimmte, konnten sie vereint werden. In einer neuen Spalte wurde im vereinten Datensatz die ursprüngliche BFAS-Version jedes Eintrags festgehalten. Alle Einträge aus BFAS-G wurden mit BFAS-Version 0 gekennzeichnet, die Einträge aus BFAS-G-R mit Version 1. Da gewisse Klienten beide Versionen des BFAS-Fragebogens ausgefüllt hatten, mussten nun erneut mehrfach vorhandene `user_id`'s entfernt werden. Zuletzt wurden die Aspekte und Dimensionen für alle Einträge neu berechnet. Dabei wurde zeilenweise für jeden Aspekt der Durchschnitt der zehn zugehörigen Features ermittelt. Für jede Dimension wurde anschließend der Durchschnitt der beiden entsprechenden Aspekte berechnet.

#### **4.2.3 Versionstrennung des Ressourcenbilderdatensatzes**

Der Datensatz des Bilderfragebogens besteht aus zwei Versionen, die nicht miteinander kompatibel sind, da bei der neuen Version neue Items hinzugefügt wurden. Der Datensatz musste in die alte und neue Version aufgeteilt werden.

Da die Werte je nach Version in unterschiedlichen Spalten abgespeichert sind, ist anhand der leeren bzw. genutzten Spalten erkennbar, zu welcher Version des Fragebogens eine Zeile gehört.

kinder-pixabay- 8_maedchen_bank.jpg	kinder-pixabay- 9_maedchen_lachen.jpg	la01_berge_blau_px.jpg	la02_bluebells_px.jpg
1.0	-1.0	NaN	NaN
NaN	NaN	1.0	1.0

Abbildung 7: Zwei Einträge im Datensatz, welche zu verschiedenen Versionen gehören.

Um zwei getrennte Datensätze zu erhalten, wurde der Ressourcenbilderdatensatz zuerst kopiert. Bei einem der zwei Datensätze wurden anschliessend alle Zeilen von Version 1 gelöscht, während beim anderen die von Version 2 gelöscht wurden. Danach wurde sichergestellt, dass jede user\_id in jedem Datensatz nur einmal vorkommt..

Die Anzahl Befragter, welche sowohl einen Ressourcenbilder- als auch einen BFAS-Fragebogen ausgefüllt hatten, war bei der alten Version des Ressourcenbilderfragebogens weitaus höher:

	Bilder alt	Bilder neu
BFAS-G	202	1
BFAS-G-R	124	42

Tabelle 3: Matrix der Anzahl Befragter pro Fragebogenversion

Deshalb wurde entschieden, diese Version weiterzuverwenden. In der Version sind folgende Kategorien von Ressourcenbildern enthalten:

- Abstrakt
- Erwachsene
- Fortbewegung
- Kinder
- Landschaft
- Nahrungsmittel
- Natur
- Relax

Für jede dieser Kategorien wurde nun reihenweise die Durchschnittlichen Werte der zugehörigen Ressourcenbilder-Features berechnet und jeweils in einer neuen Spalte abgespeichert. Die Kategorie-Durchschnitte wurden berechnet, da sie für die Datenanalyse interessant sind.

#### 4.2.4 Säuberung und Vereinigung

Unerwünschte Daten wurden aus den Dataframes entfernt. Behalten wurden die BFAS-Daten, die Ressourcenbilderdaten und gewisse persönliche Daten, ersichtlich in Tabelle 4. Da bei der Laufbahndiagnostik nicht alle persönlichen Daten Mussfelder sind enthalten gewisse Spalten leere Zellen, siehe Spalte «Komplett» in Tabelle 4. Es wurde entschieden, diese leeren Zellen ohne Modifikation beizubehalten. Nun wurden vom vereinten BFAS-Datensatz zwei Versionen generiert, eine mit und eine ohne persönliche Daten. Diese beiden Versionen wurden jeweils mit dem Ressourcenbilderdatensatz vereint, wobei die user\_id als Verknüpfungspunkt diente. Die neu generierten Dataframes wurden als BFAS\_BILD\_PERSINFO.csv und BFAS\_BILD.csv abgespeichert.

Kategorie	Datentyp	BFAS_BILD_PERSINFO.csv	BFAS_BILD.csv	Komplett
BFAS	Features			Ja
	Aspekte			Ja
	Dimensionen			Ja
	Versionsnummer			Ja
Ressourcenbilder	Features			Ja
	Kategorien			Ja
Persönliche Daten	User_id			Ja
	Geschlecht			Ja
	Altersgruppe			Ja
	Nationalität			Ja
	Berufsposition			Nein
	Berufsstatus			Nein
	Berufssektor			Nein
	Bildungsstand			Nein
	Arbeitsland			Nein
Arbeitsregion			Nein	

Tabelle 4: Überblick der Inhalte in den finalisierten Datensätzen BFAS\_BILD\_PERSINFO.csv und BFAS\_BILD.csv. Grün Gefärbte Daten wurden behalten, rot gefärbte gelöscht.

## 4.3 Datenanalyse

Ziel der Datenanalyse war es zum einen, ein besseres Gesamtbild der Daten zu bekommen und zum anderen, herauszufinden wie sehr die persönlichen Daten die Daten der Ressourcenbilder und die BFAS-Werten beeinflussen.

### 4.3.1 Stichprobe

Unsere Stichprobe enthält insgesamt 326 Befragte. 60.8% der Befragten im kombinierten Datensatz sind weiblich, die übrigen 39.2% männlich. Keine geschlechtsneutrale Person hat sowohl einen der BFAS- als auch den alten Ressourcenbilderfragebogen ausgefüllt.

Die Stichprobe ist in sechs Altersgruppen eingeteilt und deren Verteilung gleicht einer Glockenkurve. Rund 60% der Befragten ist im Alter von 30-49, mit jeweils ca. 18% im Alter von 20-29 respektive 50-59. Die Restlichen 4% sind im Alter von 10-19 oder 60-69.

25.3% der Befragten gaben keinen Berufssektor an, von den restlichen 74.7% sind die Sektoren «Gesundheits- und Sozialwesen» mit 13.3%, «Erbringung von sonstigen Dienstleistungen» mit 11.4% und «Erziehung und Unterricht» mit 8.3% die prominentesten. Fünf Sektoren wurden von lediglich 10 oder weniger Befragten ausgewählt und wurden daher bei der Analyse nicht berücksichtigt.

30.9% der Befragten gaben keine Berufsposition an, die häufigste Position war «Fachspezialist/-in» mit 36.1%, die seltenste «Obere Führungsstufe (z.B. Geschäftsleitung / CEO)» mit 6.5%. 61.7% der Befragten sind angestellt, davon 38% Teilzeit und 23.7% Vollzeit. 16.9% befinden sich in einer Ausbildung.

Insgesamt ist die Stichprobe recht klein und nicht repräsentativ für die Gesamtbevölkerung. Es ist zu bedenken, dass die Daten nicht aus einer öffentlichen Umfrage stammen, sondern von einer Plattform mit relativ geringer Reichweite. Ausserdem wurden die Fragebogen von Leuten ausgefüllt, welche aus diversen Gründen an einer Laufbahndiagnostik interessiert sind. Es könnte durchaus sein, dass diese Population ein anderes psychisches Profil hat als die Gesamtpopulation. Beispielsweise könnten gewisse Persönlichkeitsfaktoren die Chance erhöhen, dass jemand an einer Laufbahnberatung interessiert ist. Dies bedeutet nicht, dass keine wertvollen Erkenntnisse gewonnen werden können, aber diese Faktoren sollten dennoch im Hinterkopf behalten werden.

### 4.3.2 BFAS

Histogramme zeigen, dass die Verteilung der Big Five Dimensionen unter den Befragten einer Glockenkurve gleicht. Extremwerte sind durchgehend relativ selten, während der Grossteil der Werte nah an der Mitte des Wertebereichs liegt.

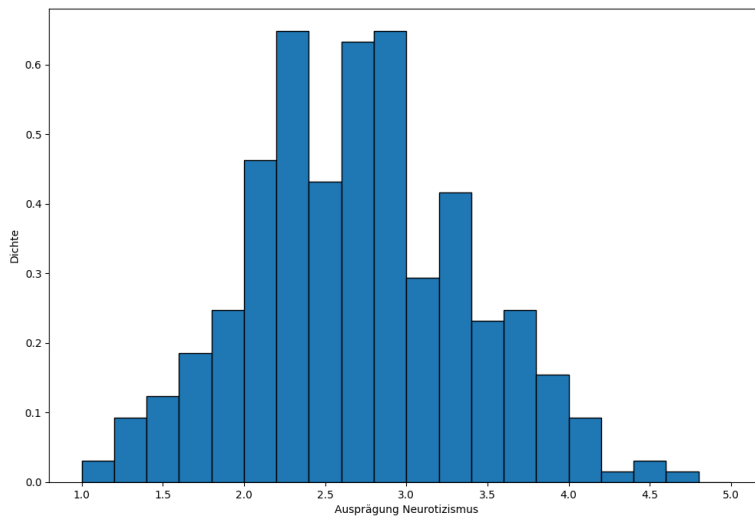


Abbildung 8: Histogramm der Verteilung der Dimension Neurotizismus.

Gewisse Dimensionen und Aspekte sind stärker gestreut, wobei die Glockenkurve schmäler und steiler ausfällt. Kennzahlen wie Mittelwert und Median, in den Plots als grünes Dreieck, respektive orange Querlinie erkennbar, sind unauffällig, Varianz ebenso. An den Daten der Big Five Dimensionen fällt auf, dass generell als positive Charaktereigenschaften empfundene Dimensionen wie «Offenheit/Intellekt» und «Verträglichkeit» die höchsten Durchschnittlichen Werte und wenig bis gar keine tiefen Werte aufweisen.

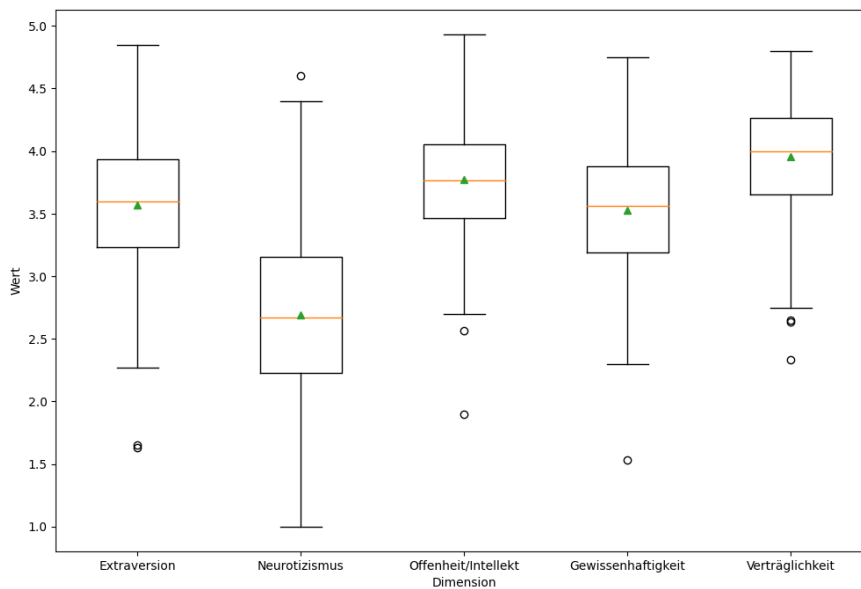


Abbildung 10: Boxplots der fünf BFAS-Dimensionen

Nur Neurotizismus, eine eher als negativ empfundene Charaktereigenschaft, weist tiefe Werte auf. Die Big Five Aspekte folgen demselben Muster wie die Dimensionen. Sie sind ähnlich verteilt und ihre Kennzahlen weichen kaum von denen der zugehörigen Dimension ab.

Eine Analyse der beiden BFAS-Versionen zeigt, dass sich einzelne Fragewerte trotz unserer Anpassungen im Durchschnitt zwischen den beiden Versionen stark unterscheiden.

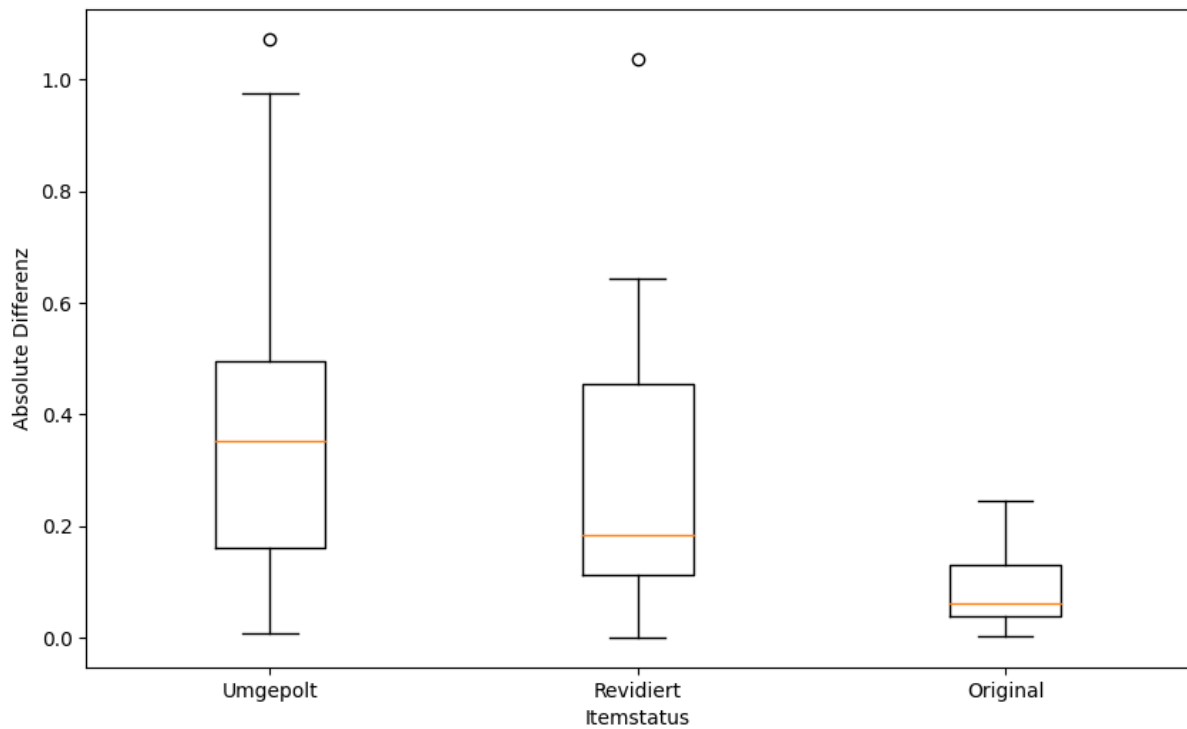


Abbildung 11: Absolute Differenz der Durchschnittswerte zwischen BFAS und BFAS-G-R nach Itemstatus.

Dies könnte mehrere Ursachen haben, beispielsweise könnte die Umformulierung der Items zu leicht unterschiedlichem Antwortverhalten bei den Befragten geführt haben, obwohl die Fragen inhaltlich gleich sind. Eine weitere Quelle für diese Unterschiede könnte die Reduktion der Skala von 7 auf 5 Werte sein. Nebst dem rein numerischen Unterschied wurde möglicherweise auch das Verhalten der Befragten beeinflusst. Die Durchschnittswerte der Dimensionen und Aspekte der beiden Versionen sind allerdings vergleichbar.

Der Einfluss von persönlichen Daten ist vor allem bei den Aspekten sichtbar. Da die Dimensionen Schnittwerte der Aspekte sind, sind die auf persönlichen Daten basierenden Unterschiede weniger sichtbar. Je nach Geschlecht unterscheiden sich die Aspekte stark, mit Unterschieden im Median von bis zu 0.53, was über 10% des gesamten Wertebereichs entspricht. Insgesamt haben weibliche Befragte im Schnitt höhere Mittel- und Medianwerte. Auffällig ist, wie hoch Median und Mittelwert der Dimension „Mitgefühl“ bei weiblichen Befragten sind und wie eng die Verteilung der Werte ist. Deren Varianz beträgt 0.22, wohingegen die der männlichen Befragten mit 0.39 fast doppelt so hoch ist.

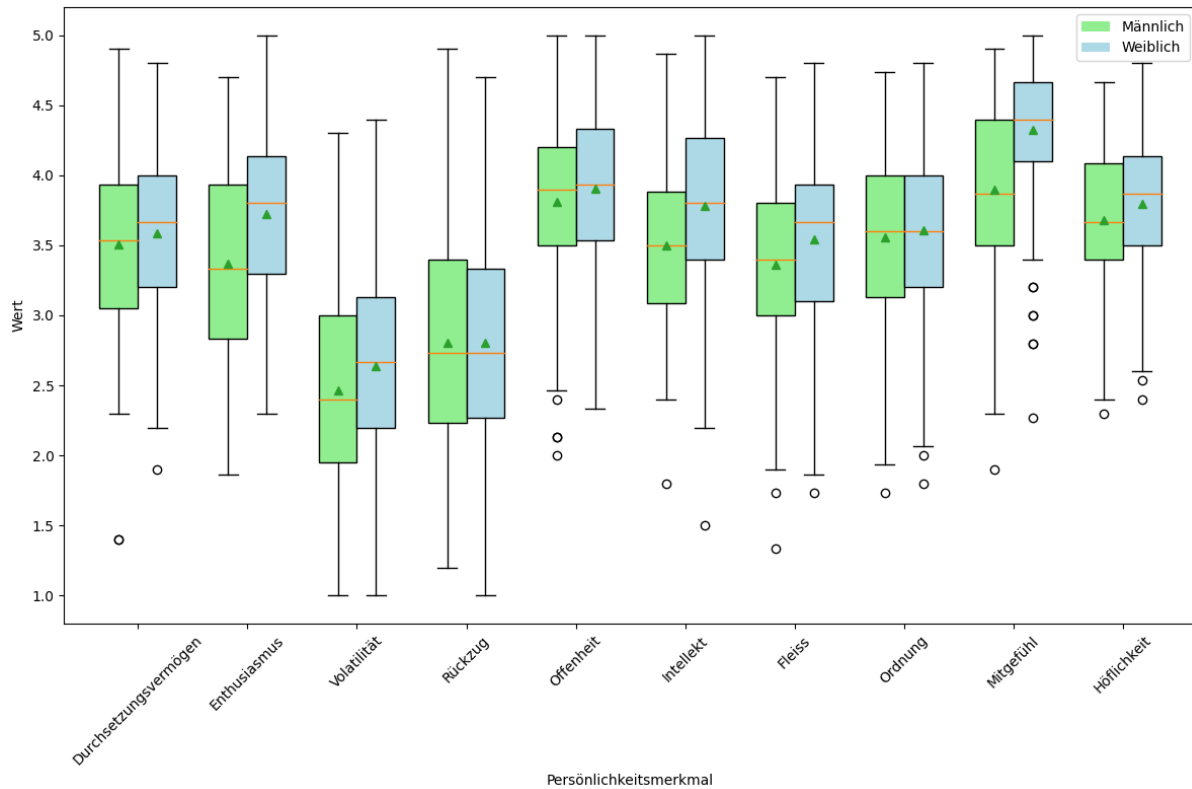


Abbildung 12: Geschlechtliche Unterschiede der BFAS-Aspekte



Um den Einfluss der weiteren persönlichen Daten auf die BFAS-Werte zu quantifizieren, wurde jeweils der Gesamt-Mittelwert und -Median der Dimensionen und Aspekte als Basis genommen. Zusätzlich wurden für die Kategorien der persönlichen Daten Median und Mittelwert berechnet. Dann wurde die Differenz zwischen den Basiswerten und den Werten der persönlichen Daten kalkuliert und untersucht.

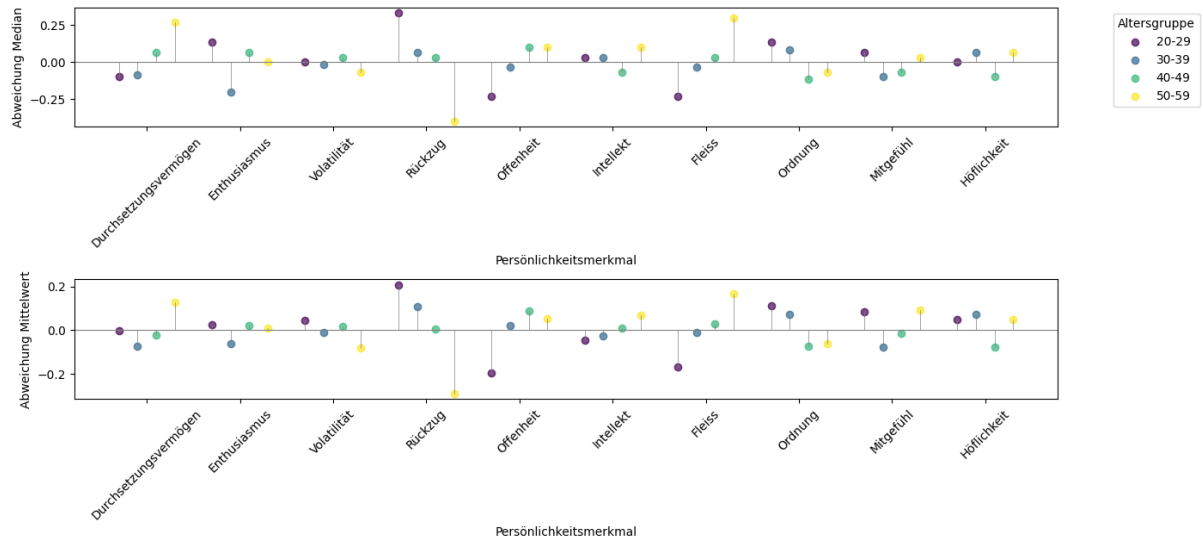


Abbildung 13: Abweichung der Durchschnittlichen BFAS-Aspekte der persönlichen Daten von den Basiswerten

Die jüngsten und ältesten Befragten tendieren bei den Aspekten dazu, in entgegengesetzte Richtungen auszuslagen, während die mittleren Altersgruppen dazwischen positioniert sind. Beispielsweise haben junge Befragte überdurchschnittliche Werte in „Rückzug“, während ältere Befragte in diesem Aspekt unterdurchschnittliche Werte aufweisen. Diese Regel hält allerdings nicht für alle Aspekte, was beispielsweise an „Mitgefühl“ und „Höflichkeit“ erkennbar ist.

Kleinere Berufssektoren mit 10 -20 Befragten weichen stark vom Basiswert ab, mit maximalen Abweichungen von über 0.6. Die grössten Gruppen, „Gesundheits- und Sozialwesen“, „Erbringung von sonstigen Dienstleistungen“, „Erziehung und Unterricht“ und „Information und Kommunikation“ weichen allerdings nur leicht von den Basiswerten ab. Berufsposition weist keine erkennbaren Muster und maximale Abweichungen von 0.35 auf. Berufsstatus besitzt auffällige Werte bei Stichprobengrößen unter 50, zum Beispiel ist der Median des Aspekts „Intellekt“ bei selbstständigen Befragten um 0.4 höher als der Basiswert.

### 4.3.3 Ressourcenbilder

Die beliebtesten Ressourcenbilder gehörten alle zur Kategorie «Landschaft»:

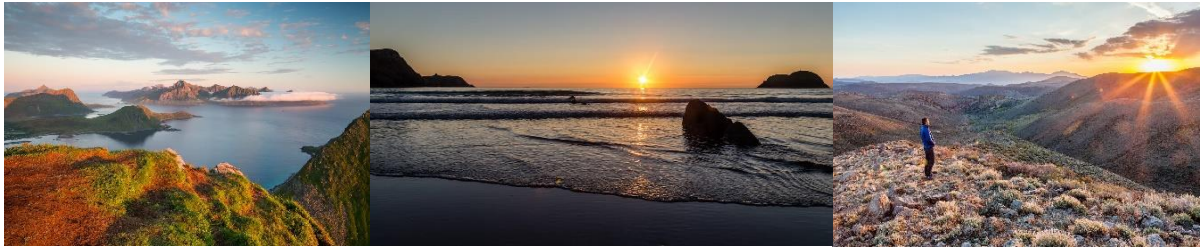


Abbildung 14: Beliebteste Ressourcenbilder, nach absteigender Beliebtheit sortiert.

Über 88% Prozent der Befragten fanden diese Bilder ansprechend.

Die unbeliebtesten Ressourcenbilder gehörten alle zur Kategorie «Fortbewegung»:



Abbildung 15: Unbeliebteste Ressourcenbilder, nach aufsteigender Beliebtheit sortiert.

Die Verteilungen der Ressourcenbilderkategorien sind sehr unterschiedlich. Die beliebteste Kategorie ist «Landschaft», sie weist den höchsten Mittelwert, 0.66 und Median, 0.73, und zugleich die tiefste Varianz, 0.08, auf. «Natur», «Relax» und «Kinder» haben die höchsten Varianzen mit 0.3. Ihre Verteilungen sind breit gestreut und rechtsschief, mit wenig negativen Werten und Vielen Positiven. Die Verteilungen von «Fortbewegung» und «Abstrakt» gleichen

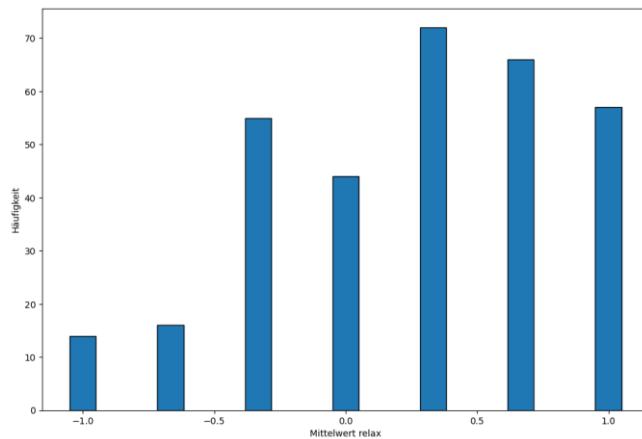


Abbildung 16: Spärliche Werteverteilung der Kategorie relax

einer Glockenkurve, mit Median und Mittelwert in der Nähe von 0. Als unbeliebteste Kategorie hat «Fortbewegung» den tiefsten Median von 0.04.

Gewisse Kategorien bestehen aus wenigen Ressourcenbildern, welche auf einer dreistufigen Skala beurteilt werden. Daraus ergibt sich eine geringe Anzahl Permutationen und folglich auch eine geringe Anzahl möglicher verschiedener Durchschnittswerte. Deswegen weisen die Verteilungen gewisser Kategorien Lücken auf, anstatt dass die Werte kontinuierlich über den Wertebereich verteilt sind. Am offensichtlichsten ist dies bei der Kategorie «Relax», welche nur 3

Ressourcenbilder beinhaltet. Die Durchschnittswerte dieser Kategorie beschränken sich also auf nur 7 verschiedene Zahlen.

Im Schnitt haben die Kategorien bei weiblichen Befragten höhere Mittelwerte. Der Median der Kategorien «Kinder» und «Natur» ist bei weiblichen Befragten mit 0.33 bzw. 0.6 signifikant höher als bei Männlichen. Die Kategorie «Fortbewegung» ist die einzige Kategorie, welche bei männlichen Befragten beliebter ist als bei Weiblichen. Zugleich ist es die einzige Kategorie, welche negative Masswerte aufweist: Der Mittelwert beträgt -0.03, der Median -0.04. Verglichen mit 0.16, bzw. 0.15 bei männlichen Befragten beträgt der Unterschied somit bei beiden Massen 0.19. Die restlichen Kategorien weisen kleinere Unterschiede auf.

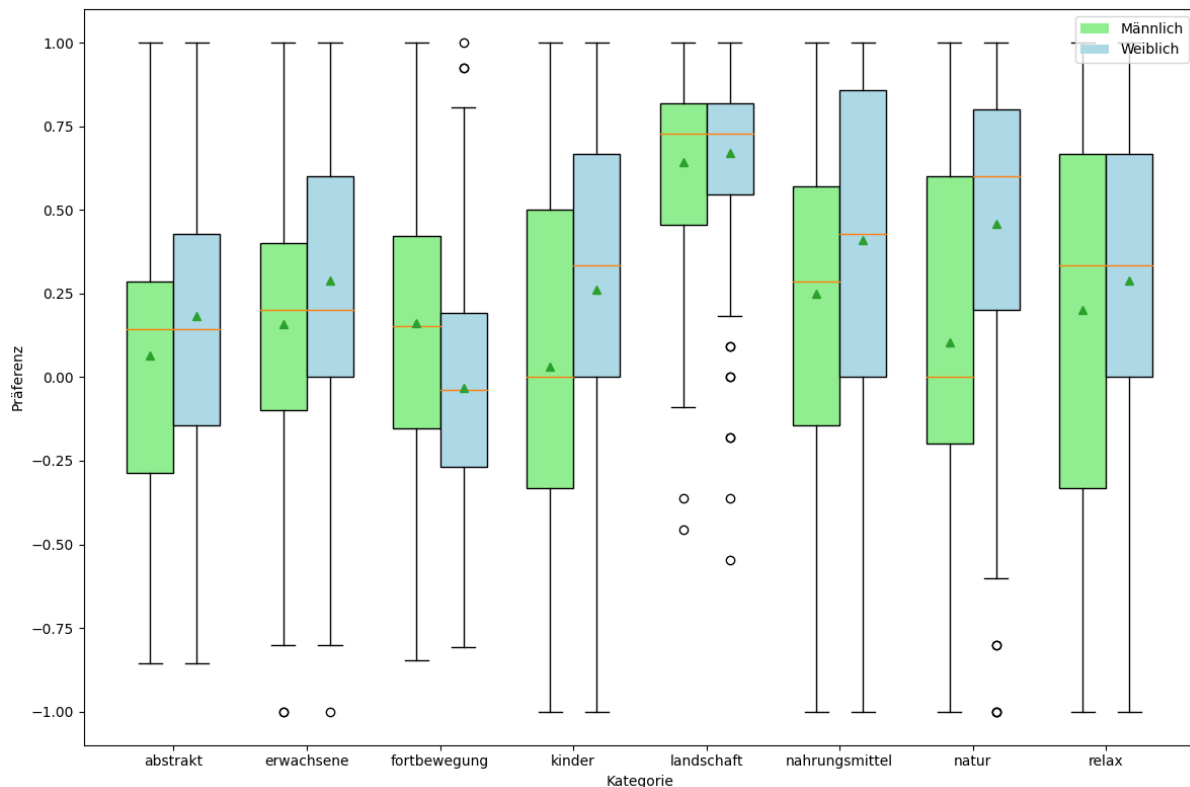


Abbildung 17: Geschlechtliche Unterschiede der Ressourcenbilderpräferenzen.

Die restlichen persönlichen Daten wurden gleich wie bei den BFAS-Daten untersucht. Die Mittelwerte und Mediane der Altersgruppe «50-59» in den Kategorien «Kinder» und «Natur» sind um fast 0.2 höher als die Basiswerte. Der Wertebereich der Ressourcenbilder ist von -1 bis 1, die Masse weichen also um fast 10% des Wertebereichs ab. Der Median in der Kategorie «Natur» ist bei den Altersgruppen «20-29» und «30-39» um 0.2 tiefer als der Basiswert.

Berufssektoren mit wenig Befragten weichen stark vom Basiswert ab, während solche mit mehr Befragten weniger stark abweichen. Bei der Berufsposition ist interessant, dass die Amplitude der Differenzen des Medians fast doppelt so gross ist wie die des Mittelwerts. Der Berufsstatus verursacht mit Amplituden von bis zu 0.4, oder 20% des Wertebereichs, die grössten Differenzen. Sie treten vor allem bei den Berufsstatus «Selbstständig» und «Arbeitslos» auf, welche nur wenig Befragte haben.

#### **4.3.4 Zusammenfassung**

Es ist kaum möglich, Aufgrund der Datenanalyse definitive Aussagen zur Korrelation zwischen den persönlichen Daten und den Fragebogendaten zu machen. Die Grösse der Stichprobe ist mit 326 Befragten klein. Bei Unterteilungen dieser Stichprobe in die verschiedenen Kategorien der persönlichen Daten beinhalten Kategorien oft weniger als 50 User. Daher ist unklar, ob die beobachteten Effekte tatsächlich existieren oder lediglich Zufall sind. Die einzige Kategorie vielversprechende Kategorie ist „Geschlecht“, mit relativ vielen Usern pro Geschlecht und signifikanten geschlechtlichen Unterschieden von bis zu 10% des Wertebereichs.

Die restlichen persönlichen Daten haben auf die BFAS-Daten weniger Einfluss als das Geschlecht. Die Differenzen zwischen den Basiswerten und denjenigen der jeweiligen Kategorie sind zumeist in der Grössenordnung 0-0.2, ca 4% des Wertebereichs. Die Altersgruppen sind interessant, weil sich die Differenzen bei gewissen Aspekten mit dem Alter linear zu verändern scheinen. Mit nur vier Altersgruppen, wovon keine über 100 Befragte beinhaltet, ist jedoch auch diese Erkenntnis nur Spekulation.

Die Ressourcenbilder scheinen insgesamt mehr von den persönlichen Daten beeinflusst zu sein.

## 5. Clustering

Es wurden beim Clustering mehrere Ziele verfolgt. Erstens sollten die Daten explorativ analysiert werden, um Muster oder Anomalien aufzuzeigen. Zweitens sollte Mittels Cluster Correspondence Analysis die Beziehung zwischen den Ressourcenbilder- und BFAS-Daten erkundet werden.

### 5.1 Methoden

#### 5.1.1 Clustering Algorithmen

Für diese Arbeit wurden zwei Clustering-Algorithmen gebraucht. Diese Algorithmen wurden aufgrund ihrer relativen Einfachheit und Geschwindigkeit gewählt.

##### *K-Means*

K-Means hat einen Parameter, eine ganze Zahl  $k > 1$ . K-Means teilt die Daten in  $k$  Cluster. Der Algorithmus beginnt, indem  $k$  Clusterzentroide zufällig im Raum verteilt werden. Daraufhin werden abwechselungsweise die Datenpunkte eingeteilt, je nachdem zu welchem Zentroid sie gehören und die Zentroide so verschoben, dass der Abstand zu den zugehörigen Datenpunkten minimiert wird. Der Abstand wird mit der euklidischen Distanz gemessen.

Bei vielen Dimensionen wächst die euklidische Distanz jedoch stark an, sodass Datenpunkte weiter voneinander entfernt erscheinen als in niedrigdimensionalen Räumen. In hochdimensionalen Räumen verliert die euklidische Distanz an Aussagekraft, da die Unterschiede zwischen den kürzesten und längsten Distanzen relativ gering sind. Dadurch wird es schwieriger, klare Cluster zu identifizieren, und K-Means funktioniert in hochdimensionalen Räumen nicht gut.

##### *DBSCAN*

DBSCAN hat zwei Parameter: die ganze Zahl  $\text{minPts} > 1$  und die Gleitkommazahl  $\text{epsilon} > 0$ . DBSCAN gruppiert Datenpunkte, indem für jeden Datenpunkt die Anzahl Nachbarpunkte innerhalb des Radius  $\text{epsilon}$  ermittelt wird. Ist diese Anzahl grösser oder gleich  $\text{minPts}$ , so werden die Punkte zusammen gruppiert. Hat ein Punkt keine oder zu wenig Nachbarn, wird er als Ausreisser klassifiziert.

Das Clustering von DBSCAN basiert nicht auf Abständen, sondern auf der Dichte der Daten. Dadurch performt DBSCAN in hochdimensionalen Räumen besser als K-Means. Ausserdem macht DBSCAN keine Annahmen über die Anzahl und Form der Cluster und ist gegen Ausreisser resistent. Der Nachteil an DBSCAN ist, dass es schwieriger ist, für zwei Parameter die optimalen Werte zu ermitteln.

#### 5.1.2 Dimensionality Reduction

Dimensionality Reduction wurde aus zwei Gründen gemacht: Erstens funktionieren Clustering-Algorithmen wie K-Means auf Datensätzen mit vielen Dimensionen nicht gut. Wenn beispielsweise die 100 BFAS-Features auf Cluster untersucht werden sollen, würde K-Means keine guten Ergebnisse liefern. Zweitens ist die Visualisierung von hochdimensionalen Daten nicht möglich. Deshalb wurden zwei Verfahren genutzt, um die Dimensionalität der Daten zu reduzieren.

## PCA

PCA generiert orthogonale Linearkombinationen der ursprünglichen Dimensionen des Datensatzes, um die maximale Varianz zu erklären. Jede neue Komponente wird so berechnet, dass sie die grösstmögliche verbleibende Varianz erklärt und orthogonal zu den vorherigen Komponenten bleibt. Das Ergebnis ist ein transformierter Datensatz mit folgenden Eigenschaften:

- Die erste Komponente erklärt die meiste Varianz.
- Jede nachfolgende Komponente erklärt den maximal möglichen Anteil der verbleibenden Varianz.
- Da die gesamte Varianz des ursprünglichen Datensatzes erhalten bleibt, geht keine Information verloren.

PCA verändert die zugrundeliegenden Daten nicht, sondern projiziert sie auf ein neues Koordinatensystem. Daraufhin können die wichtigsten Dimensionen für Clustering und Visualisierung ausgewählt werden. Dabei wird der Informationsverlust minimiert und die Performance von Clustering-Algorithmen verbessert.

## T-SNE

T-SNE berechnet Wahrscheinlichkeiten, die die Ähnlichkeit von Datenpunkten in einem hochdimensionalen Raum darstellen, wobei nahe beieinander liegende Punkte hohe Wahrscheinlichkeiten erhalten. Dann versucht der Algorithmus die Datenpunkte auf einen tiefdimensionalen Raum zu projizieren, wo die Ähnlichkeit erneut berechnet wird. Diese Projektion wird optimiert, indem die Divergenz zu der originalen hochdimensionalen Projektion minimiert wird. T-SNE ist gut darin, lokale Strukturen in hochdimensionalen Daten auch in der tiefdimensionalen Projektion widerzuspiegeln. Die globale Verteilung der hochdimensionalen Daten wird allerdings in der tiefdimensionalen Projektion nicht gut sichtbar. [18]. T-SNE eignet sich daher sehr, um herauszufinden, welche Datenpunkte im Originaldatensatz nahe beieinander liegen.

### 5.1.3 Clusteroptimierung

#### *Silhouette Score:*

Der Silhouette Score misst die Qualität eines Clustering-Ergebnisses auf einer Skala von -1 bis 1. Dafür wird die Nähe der Datenpunkte zu den Nachbarsclustern gemessen. Sind die Nachbarscluster weit weg, ist der Score für den Datenpunkt nahe bei 1. Wenn der Datenpunkt an der Grenze zwischen Clustern liegt, ist der Silhouette Score nahe bei 0. Ein Silhouette Score in der Nähe von -1 deutet darauf hin, dass ein Punkt im falschen Cluster eingeteilt wurde.

#### *Inertia:*

Inertia ist ein Maß für die Summe der quadrierten Abstände zwischen den Datenpunkten und ihren nächstgelegenen Clusterzentroiden. Es gibt die Kompaktheit der Cluster an, wobei geringere Werte auf dichtere und besser definierte Cluster hinweisen. Ein niedriger Inertia-Wert ist wünschenswert, bedeutet jedoch nicht immer eine bessere Clustering-Qualität, da viele Cluster automatisch kleine Inertia-Werte bedeuten.

Wenn Inertia in Abhängigkeit der Zentroidzahl kalkuliert und visualisiert wird, kann man auf einen sogenannten «Elbow Point» stossen. Der Elbow Point ist ein Knick in einer Inertia-Kurve, bei dem die Addition weiterer Cluster keine starke Reduktion von Inertia mitbringt. Daher ist der Elbow Point ein guter Indikator für die optimale Anzahl Cluster.

### 5.1.4 Cluster Correspondence Analysis

Cluster Correspondence Analysis (CCA) untersucht die Beziehung zwischen zwei unterschiedlichen Clusterings. Es wird verwendet, um zu bestimmen, wie gut die Clusterstrukturen zweier Datensätze übereinstimmen. In dieser Arbeit werden die besten Clusterings der Ressourcenbilder und der BFAS-Daten mittels CCA verglichen.

#### Adjusted Rand Index:

Der Adjusted Rand Index (ARI) ist ein Maß für die Übereinstimmung zwischen zwei Clusterings, welches die Anzahl der Übereinstimmungen und Nichtübereinstimmungen von Datenpunkten berücksichtigt, wobei Zufallseffekte herausgerechnet werden. Der ARI-Wert liegt zwischen -1 und 1, wobei 1 eine perfekte Übereinstimmung und 0 eine Übereinstimmung auf Zufallsniveau bedeutet.

#### Normalized Mutual Information:

Normalized Mutual Information (NMI) quantifiziert die normalisierte geteilte Information zweier Clusterings in einem Wert zwischen 0 und 1. Ein Wert von 1 bedeutet perfekte Übereinstimmung mit maximaler geteilter Information, während ein Wert von 0 bedeutet, dass die Clusterings keine Information teilen. NMI berücksichtigt sowohl die Entropie der Clusterings als auch die gemeinsame Information.

## 5.2 Ergebnisse

Die Beziehungen zwischen den BFAS-Aspekten wurden untersucht, indem die Aspekte paarweise miteinander geplottet wurden. Dabei wurde festgestellt, dass Aspekte derselben Dimension oftmals positiv korrelieren. Besonders ausgeprägt ist dies bei den Aspekten «Rückzug» und «Volatilität». Auch gewisse, scheinbar nicht zusammenhängende Aspekte weisen positive Korrelationen auf, wie «Fleiss» und «Durchsetzungsvermögen», oder «Mitgefühl» und «Enthusiasmus». Es wurden auch ein paar negative Korrelationen festgestellt, beispielsweise korreliert «Fleiss» sowohl mit «Rückzug» als auch mit «Volatilität» negativ. «Fleiss» scheint also mit der Dimension «Neurotizismus» eine inverse Beziehung zu haben.

Die Dimensionalität der BFAS- und Ressourcenbilderdatensätze wurde mittels PCA und t-SNE reduziert, damit die Datensätze visualisiert und geclustert werden konnten. Beim BFAS-

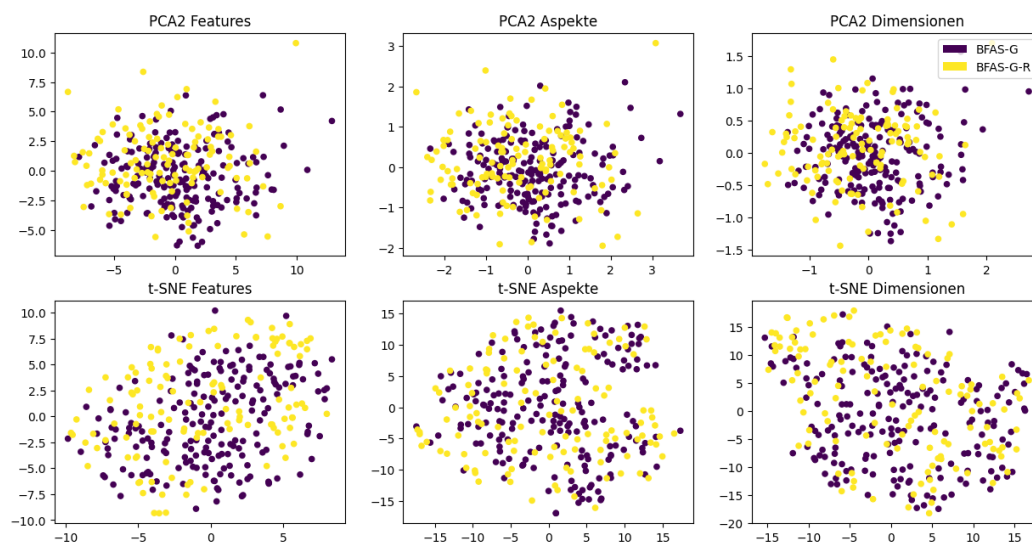


Abbildung 18: Dimensional reduzierte BFAS-Daten, nach Version gefärbt.

Datensatz wurden sowohl die 100 Features als auch die 10 Aspekte und 5 Dimensionen mittels PCA auf 2 Dimensionen reduziert. Dann wurde untersucht, ob die Unterschiede zwischen den BFAS-Versionen sichtbar sind. Die resultierende Grafik weist keine von Auge erkennbaren Cluster auf. Ausserdem sind die beiden BFAS-Versionen zumindest in den ersten zwei PCA-Dimensionen gleichmässig gestreut und überlappen sich.

Diese Analyse wurde mit t-SNE wiederholt, wobei beobachtet wurde, dass sich die Datenpunkte der zwei BFAS-Versionen hier im Vergleich zu PCA weniger überlappen. T-SNE behält lokale Strukturen hochdimensionaler Daten bei. Die bisweilen klare Trennung der zwei BFAS-Versionen deutet darauf hin, dass in den BFAS-Daten kleine Gruppen ähnlicher Datenpunkte existieren, welche zumeist aus der gleichen BFAS-Version stammen. Eine mögliche Ursache dafür ist, dass die Werte numerisch verschieden sind, da BFAS-G auf einer 7-stufigen Skala basiert, welche auf einen Wertebereich von 1 bis 5 skaliert wurde. Zum anderen wurde ein grosser Teil der Fragen revidiert, was zu Unterschieden in den Durchschnittswerten zwischen den Fragebögen führt.

Die Performance von PCA auf den BFAS-Features, Aspekten und Dimensionen wurde untersucht, um herauszufinden, wie gut die zweidimensionale Projektion den ursprünglichen Datensatz repräsentiert. Dazu wurde die kumulierte erklärte Varianz pro PCA-Komponente kalkuliert. Die ersten beiden PCA-Komponenten erklären 22.5% der Varianz aller Features. Die ersten beiden PCA-Komponenten der Aspekte erklären 50% der Varianz. Das bedeutet einen Informationsverlust von 50%, wenn die Aspekte mittels PCA auf zwei Komponenten reduziert werden. Die ersten beiden PCA-Komponenten der Dimensionen erklären 62% der Varianz. Um 90% der Varianz zu repräsentieren, würden bei den Features 56, bei den Aspekten sieben und bei den Dimensionen vier PCA-Komponenten benötigt. Diese Ergebnisse deuten darauf hin, dass die Varianz der BFAS-Daten relativ gleichmässig verteilt ist. Ausserdem reichen die ersten zwei PCA-Komponenten der Features für eine Repräsentation der Originaldaten nicht aus.

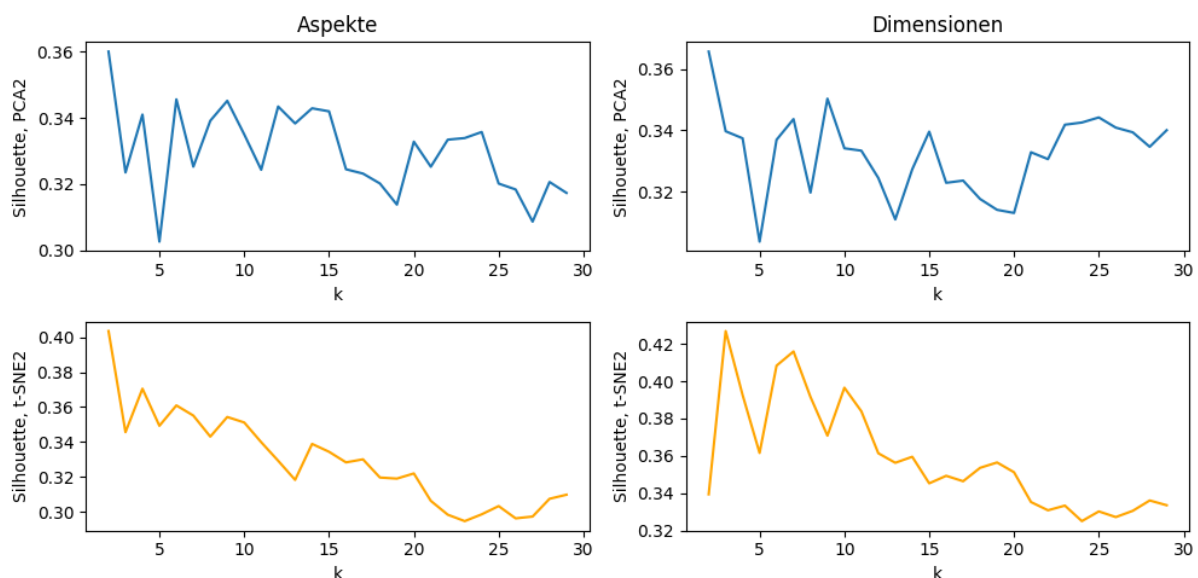


Abbildung 19: Verlauf der Silhouette Scores in Abhängigkeit von k.

Die Ressourcenbilderdaten wurden ebenfalls mit PCA und t-SNE dimensional reduziert und visualisiert, weisen jedoch keine interessanten Muster auf. Die erklärte Varianz der ersten paar PCA-Komponenten ist tief. Zwei Komponenten erklären beispielsweise weniger als 30% der Varianz. Die darauf basierenden zweidimensionalen Visualisierungen sind also keine gute Repräsentation des Gesamtdatensatzes. Bei der Frage, ob Beziehungen zwischen den



Ressourcenbilderdaten und den BFAS-Aspekten und -Dimensionen existieren, sind die BFAS-Features belanglos. Deswegen werden die BFAS-Features hier nicht weiter erwähnt.

Mithilfe von Silhouette Scores und Inertia wurde ein optimales  $k$  für ein  $k$ -Means Clustering auf den zweidimensionalen PCA und t-SNE Daten gesucht. Es wurden  $k$ -Werte von 2 bis 30 untersucht, sichtbar in Abbildung 19. Die Silhouette Scores befinden sich durchgehend im positiven Bereich, schwanken allerdings stark. Die besten Silhouette Scores wurden auf den t-SNE-Daten generiert, mit  $k=2$  für die Aspekte und  $k=3$  für die Dimensionen. Die Inertia-Werte weisen keinen klaren «Elbow-Point» auf. Clustering auf den Originalen Aspekt- und Dimensionsdaten generierte schlechtere Silhouette Scores. Mittels Silhouette Plots wurden die vielversprechendsten Clusterzahlen untersucht. Die Verteilungen der Silhouette Scores sind breit und fallen linear ab. Die Cluster enthalten wenige negative Silhouette Scores.

Die Ressourcenbilder wurden identisch analysiert. Der Maximale Silhouette Score der PCA Dimensionalitätsreduktion liegt mit 0.36 leicht unter dem t-SNE Maximum von 0.42. Beide Maxima wurden mit  $k=2$  erzielt. Erneut konnte kein klarer Elbow Point festgestellt werden. Die Silhouette Plots weisen maximale Silhouetten von über 0.6 und keine negativen Werte auf.

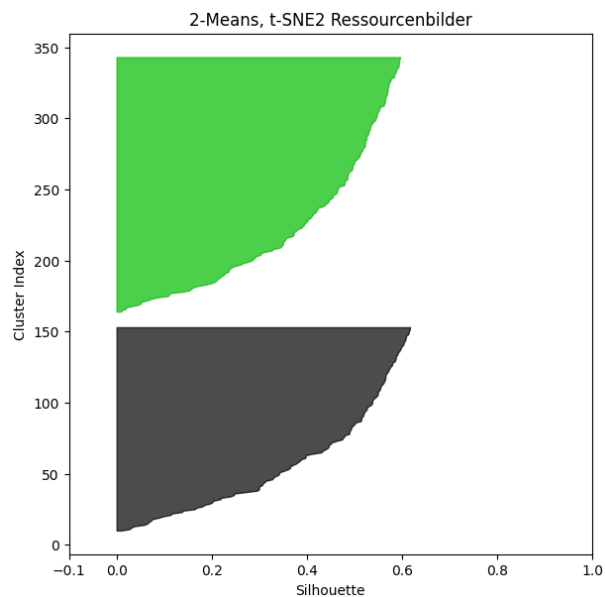


Abbildung 20: Silhouette Plot der Ressourcenbilder

Die optimalen Parameter für DBSCAN wurden ermittelt, indem minPts und Epsilon variiert und die Silhouette maximiert wurde. DBSCAN auf zwei PCA-Komponenten der BFAS-Aspekte lieferte den besten Silhouette Score von 0.61 mit minPts = 2 und epsilon = 1.15. Das beste Ergebnis der BFAS-Dimensionen war ein Silhouette Score von 0.57 mit minSamples = 4 und epsilon = 1.11. Dieses Ergebnis wurde ebenfalls mit DBSCAN auf zwei PCA-Komponenten erreicht. DBSCAN auf zweidimensionalen t-SNE Dataframes erzielte konsistent tiefere Werte als DBSCAN auf den Originaldaten. Der Bestwert der Ressourcenbilder wird ebenfalls mit DBSCAN auf zwei PCA-Komponenten erzielt. Der Silhouette Score beträgt 0.45 und wird mit minPts = 5 und epsilon = 0.64 erreicht. Insgesamt wurden sowohl für die BFAS- als auch für die Ressourcenbilderdaten die besten Ergebnisse mit DBSCAN erreicht. Eine weitere Analyse der DBSCAN Cluster mit maximalem Silhouette Score zeigt allerdings, dass die Mehrheit der Datenpunkte als Ausreisser klassifiziert sind. Die Generierten Cluster sind sehr klein, mit jeweils nur wenigen Datenpunkten. Der beste insgesamt erreichte Silhouette Score wurde von DBSCAN auf t-SNE reduzierten BFAS-Aspekten generiert. Die dazugehörigen Cluster Labels zeigen insgesamt 74 Cluster, wovon der Grossteil nur zwei Datenpunkte beinhaltet.

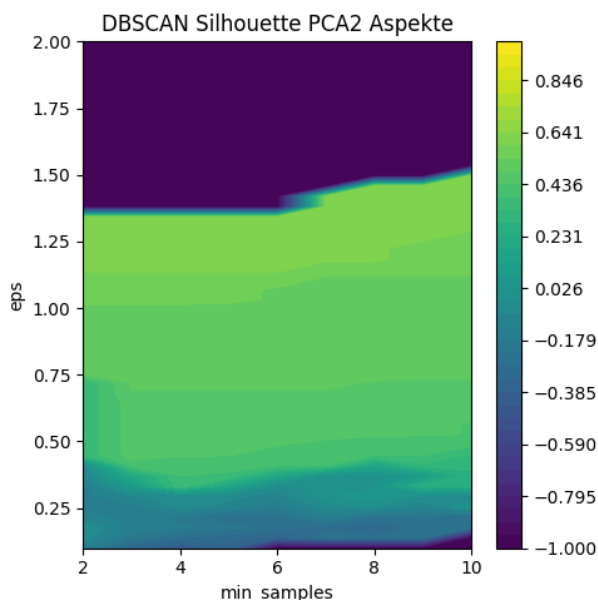


Abbildung 21: Contourplot der DBSCAN Optimierung auf zweidimensionaler PCA-Projektion der Aspekte

Die Cluster Correspondence Analysis ergibt durchgehend sehr tiefe Werte für den ARI. Auffälligerweise ist aber der NMI der Clusterings der BFAS-Aspekte und BFAS-Dimensionen deutlich höher. Diese Werte sind im Kontext der zuvor erwähnten Clusterings nicht überraschend.

Daten	Algorithmus	Dimension	ARI	NMI
Ressourcenbilder	DBSCAN	PCA 2D	0.001	0.026
BFAS-Aspekte	DBSCAN	PCA 2D		
Ressourcenbilder	DBSCAN	PCA 2D	0.014	0.024
BFAS-Dimensionen	DBSCAN	PCA 2D		
BFAS-Aspekte	DBSCAN	PCA 2D	0.025	0.356
BFAS-Dimensionen	DBSCAN	PCA 2D		

Zusammenfassend lässt sich sagen, dass mit den besprochenen Methoden weder in den BFAS- noch in den Ressourcenbilderdaten zufriedenstellende Cluster gefunden werden konnten. Die visuelle Analyse mittels Dimensionalitätsreduktion liess dies bereits vermuten. Die von

DBSCAN generierten Clusterings waren unbrauchbar, obwohl der Silhouette Score sehr vielversprechend war.

## 6. Prediction Models

### 6.1 Vorgehensweise

Nach dem Scheitern des Versuchs des Clustering, wurde mit Hilfe von Prediction Models versucht, herauszufinden ob BFAS-Werte aufgrund der Ressourcenbilderdaten vorausgesagt werden können. Dazu wurden folgende Testszenarien definiert:

Testcase 1	Predict BFAS Categories from Ressourcenbilder data
Testcase 2	Predict BFAS subcategories from Ressourcenbilder data
Testcase 3	Predict BFAS Categories from Ressourcenbilder data & personal information
Testcase 4	Predict BFAS subcategories from Ressourcenbilder data & personal information

Testcase 1 und 2 beinhalteten als Input alle Bildpräferenzen der Probanden als Werte auf einer Skala von -1, 0 oder 2. Die Gruppierungen der positiv bewerteten Bilder und der Text wurden aussenvor gelassen. Testcases 3 und 4 beinhalten ausserdem die folgenden persönlichen Features:

- Geschlecht
- Alter
- Berufsposition, beispielsweise «Sachbearbeiter/-in»
- Berufsstatus, beispielsweise «Teilzeitangestellt»
- Berufssektor, beispielsweise «Gesundheits- und Sozialwesen»
- Land
- Arbeitskanton
- Bildungsstand

Diese persönlichen Features wurden durch One-Hot-Encoding ungewandelt, damit die Modelle funktionieren. Die Y-Werte, also die vorherzusagenden Werte sind jeweils die neu berechneten Scores pro Person der jeweiligen Kategorien.

Das ganze Modell wurde mit den folgenden Algorithmen ausgewertet:

- Linear Regression (aus der Library sklearn)
- xGBoost Random Forest Regressor (Extreme gradient boosting, aus der Library xgboost)
- Random Forest Regressor (aus der Library sklearn)

Im Versuch wurde der Mittelwert als Baseline benutzt. Der Mittelwert wurde ausschließlich aus den Trainingsdaten berechnet und anschließend mit derselben Validierungsmethode validiert. Für die Validierung wählten wir den mittleren quadratischen Fehler (MSE, Mean-Squared-Error).

Da es sich um relativ wenige Datenpunkte handelt, wurde sowohl mit n-fold Cross-Validation (n=5) als auch mit einer Leave-One-Out Cross-Validation gearbeitet, bei der die Test-Splits genau ein Feature sind.

Im Clustering hat sich gezeigt, dass Dimensionsreduktion zum Teil zu besseren Ergebnissen führen kann. Deshalb wurde der Versuch so erweitert, dass die Inputdaten mit PCA in der Dimensionalität reduziert werden. Danach wurde mit inkrementell mehr PCA-Variablen der

gleiche Versuch laufen gelassen. Die maximale Anzahl der PCA-Variablen wurde auf 5 gesetzt. Aus all diesen unterschiedlichen Anzahlen der PCA-Variablen wurde dann der tiefste Wert des MSE mit der unterschiedlichen Anzahl PCA-Features als schlussendliches Ergebnis gewertet.

Zusammenfassend ergeben sich also folgende Testszenarien:

- 4 Testcases
- Mit je 2 unterschiedlichen Cross-Validations
- Mit und ohne Dimensionalitätsreduktion von PCA

In den nachfolgenden Grafiken sind jeweils die Dimensionen/Aspekte gruppiert zu sehen, die verschiedenen Farben stellen die verschiedenen verwendeten Algorithmen dar. Die Baseline, der Mean ist mit gelb dargestellt.

## 6.2 Ergebnisse

In den Ergebnissen ist ersichtlich, dass sich die verschiedenen Dimensionen unterschiedlich gut vorhersagen lassen. Dabei stechen besonders die Kategorien Offenheit/Intellekt und Verträglichkeit heraus, die die tiefsten MSE erzielen.

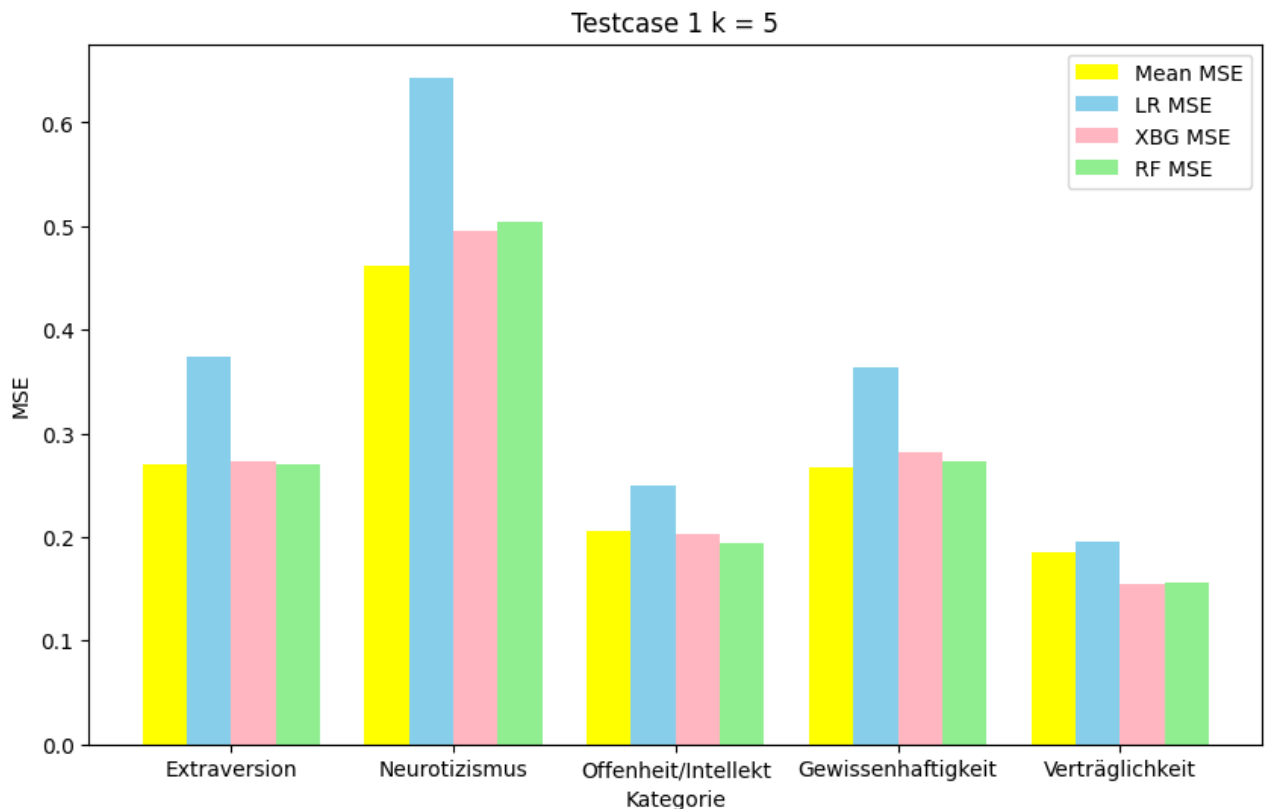


Abbildung 22: Testcase 1, k = 5 Crossvalidation, MSE

Es fällt auf, dass die Lineare Regression am schlechtesten funktioniert, während die Random Forest Modelle bessere Ergebnisse liefern. Ebenfalls auffällig ist, dass der MSE des Means relativ tief ist, und nur selten von den Modellen unterschritten wurden.

Vergleicht man jetzt Testcase 1 mit Testcase 2, also die BFAS-Dimensionen mit den -Aspekten, dann lässt sich ein Zusammenhang erkennen. Die Dimensionen mit tiefen MSE zeigen auch in deren Aspekten tiefe Werte und umgekehrt. Die MSE sind jedoch etwas höher bei den

Dimensionen, gegenüber den Aspekten. Ein möglicher Grund dafür könnte sein, dass die Dimensionen jeweils die Durchschnitte der beiden zugehörigen Aspekte sind.

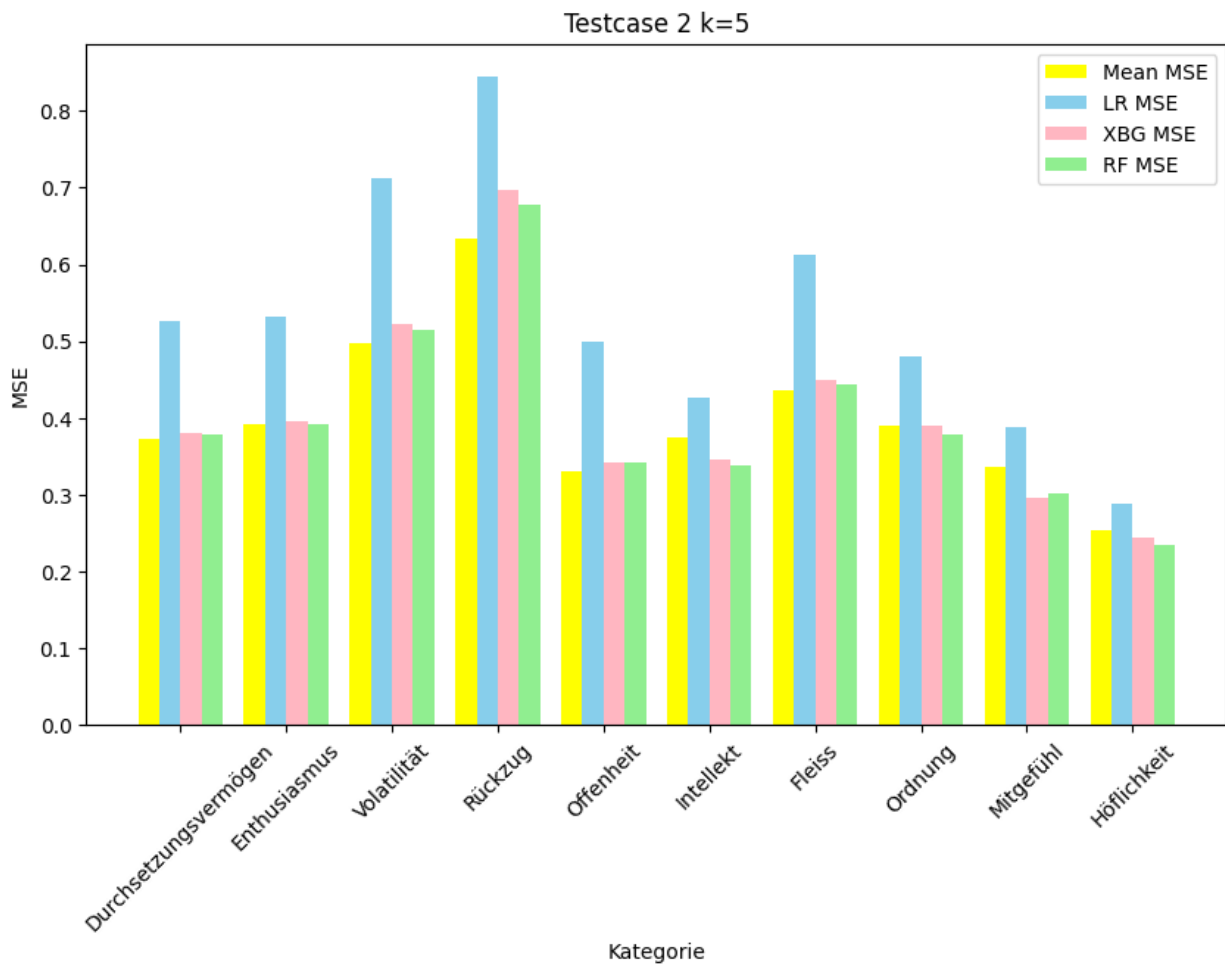


Abbildung 23: Testcase 2,  $k=5$  Crossvalidation, MSE

Als nächstes wurde geschaut, ob die Persönlichen Daten einen Einfluss auf die Vorhersagen haben. Wenn man Abbildung 20 mit Abbildung 22 vergleicht, bei denen der einzige Unterschied das Hinzufügen der persönlichen Daten sind, sieht man, dass sich nicht viel ändert. Das Einzige, was auffällt, ist dass die Lineare Regression schlechter wird.

Ein Weg, die Lineare Regression zu verbessern, war die Leave-One-Out Crossvalidation. Die MSE verbessern sich in der Differenz zum Mean um zirka 50%.

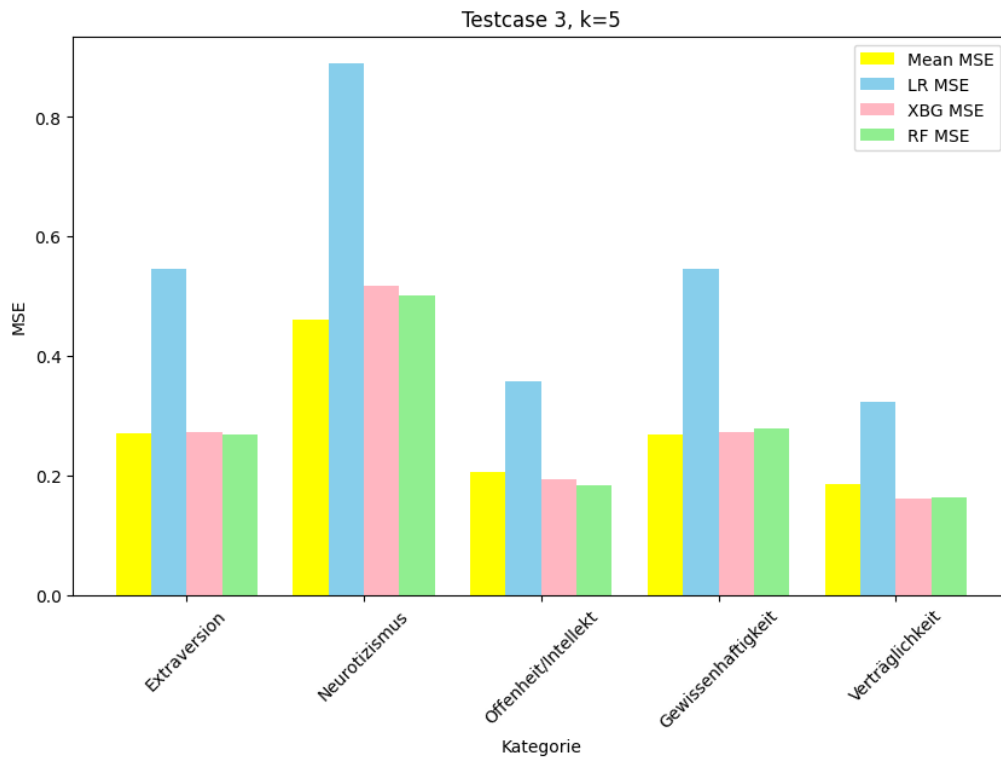


Abbildung 24: Testcase 3, k =5 Crossvalidation, MSE

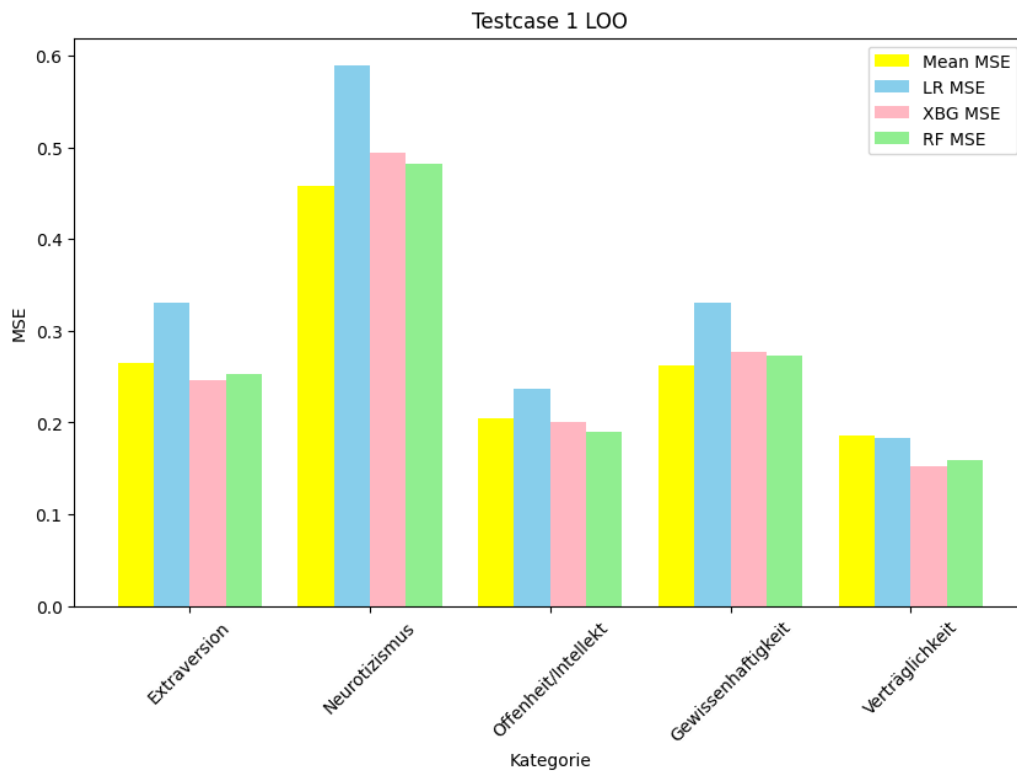


Abbildung 25: Testcase 1, Leave-One-Out Crossvalidation, MSE

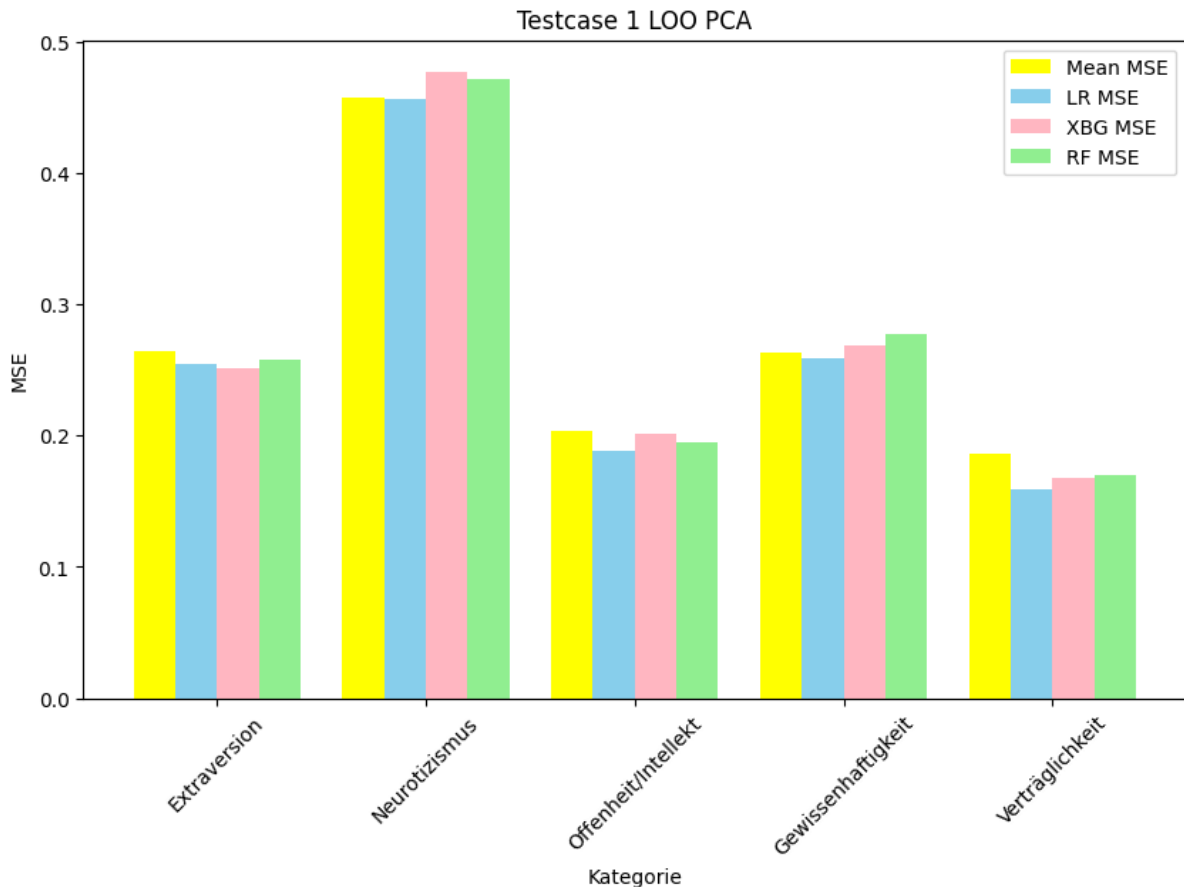


Abbildung 26: Testcase 1, Leave-One-Out Crossvalidation, PCA, MSE

Spannend ist der Einfluss der Dimensionalitätsreduktion durch PCA. Die Lineare Regression verbessert sich deutlich, während die Random Forests ungefähr die gleichen Werte erzielen. Auffällig ist, dass sich alle 4 MSE im Wert nähern, wie in Abbildung 24 sichtbar ist. Diese Grafik zeigt den Unterschied zwischen dem MSE des Modells und der Baseline. Alle vorhergesagten MSE liegen extrem nahe am Mean.

Wenn man nun alle Testcases anschaut, so kann man zusammenfassend sagen:

- Lineare Regression liefert die schlechtesten Ergebnisse
- Kategorie Offenheit/Intellekt und Verträglichkeit (O & A) lassen sich am besten vorhersagen und liefern die besten Ergebnisse
- Leave-One-Out Crossvalidation liefert minimal verbesserte Ergebnisse
- PCA Dimensionalitätsreduktion verbessert die Lineare Regression und verschlechtert die Random Forest Modelle
- Die Persönlichen Informationen, die in Testcases 3&4 einbezogen wurden, haben einen negativen Einfluss auf unsere Modelle
- Es funktioniert besser die BFAS Dimensionen hervorzusagen als die Aspekte
- Die Baseline allein liefert schon gute Ergebnisse

Die restlichen Grafiken für alle Testcases befinden sich im Anhang.

### 6.3 Weitere Analysen

Wenn man nur die Ergebnisse der Modelle anschaut, sehen die Resultate relativ gut aus. Jedoch stellte sich durchgehend die Frage, weshalb der MSE vom Mean so gut ist. Je nach Kategorie erzielt der Mean einen MSE von zwischen 0.18 und 0.46, was einer Abweichung von 4.5% bis 11.5% entspricht auf der Skala von 1 bis 5.

Um das Ganze etwas genauer zu untersuchen, wurde der ganze Versuch noch einmal neu aufgebaut und genauer analysiert.

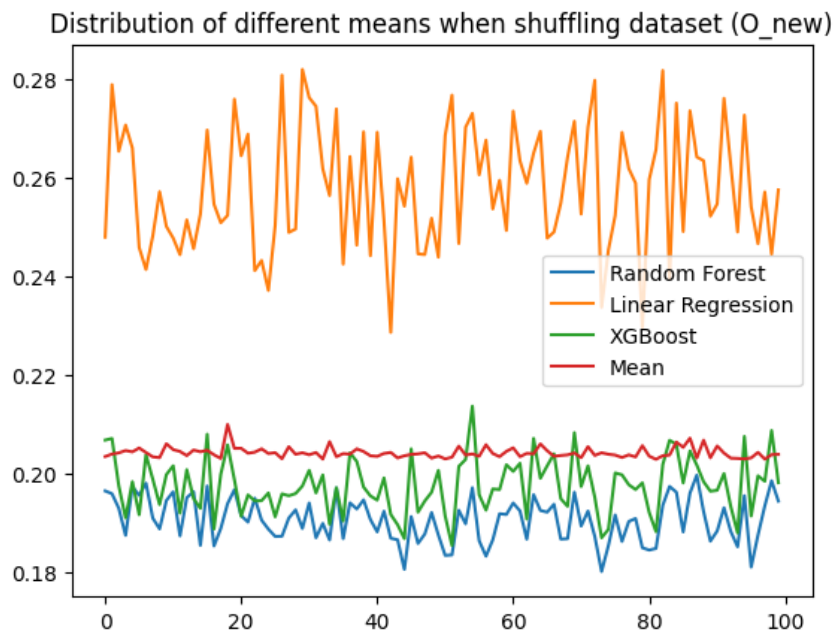


Abbildung 27: Verteilung von der MSE von 100 verschiedenen Durchläufen der Vorhersagen

Der oben gezeigte Graph soll zeigen, dass die Ergebnisse konstant waren und nicht durch Zufall der Crossvalidation. Man sieht, dass die Lineare Regression auch hier konstant schlechter ist, wie der Mean und in diesem Fall die beiden Random Forest Modelle fast ausschliesslich bessere Ergebnisse haben wie die Baseline.

Eine weitere Analyse, die vollzogen wurde, war die Betrachtung der Reste. Hier offenbarten sich vor allem bei der Linearen Regression einige Probleme.



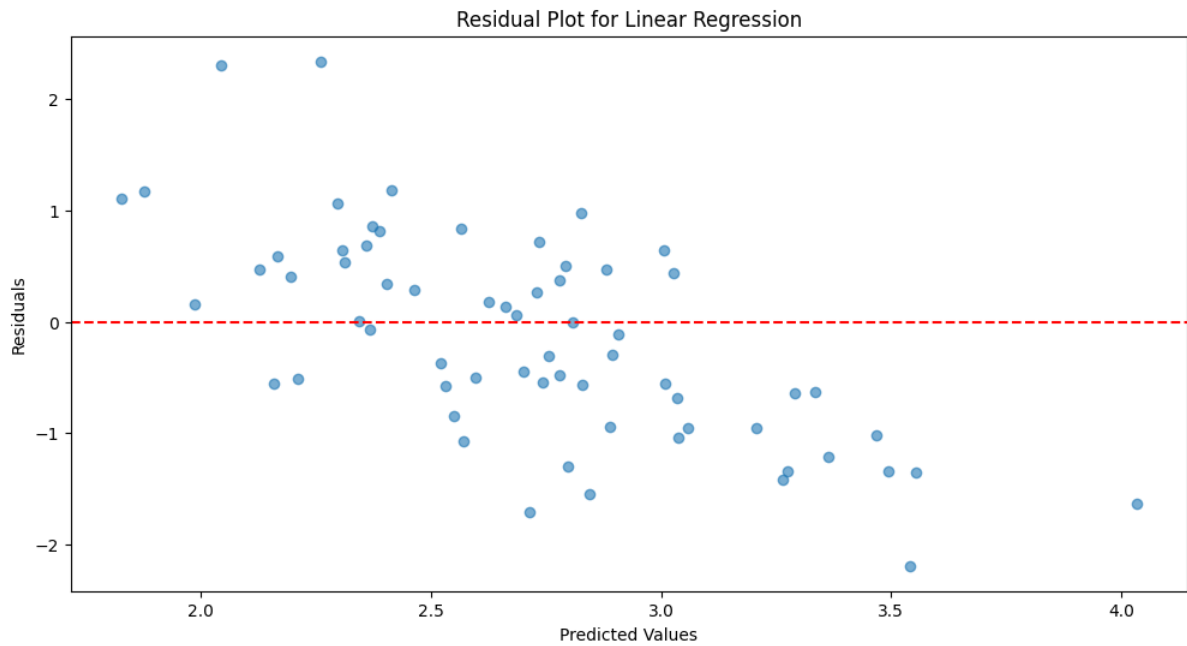


Abbildung 28: Restplot der Ergebnisse der Vorhersage der Kategorie Neurotizismus

Wie man sehen kann, konnte in den Daten keine Linearität erkannt werden. Die Fehler sind für eine Skala von 1 bis 5 zwischen + 2 und -2 extrem gross. Zwar orientiert sich der Grossteil der um die Mittellinie, aber sehr präzise sind nur wenige Vorhersagen. Schaut man sich die Reste von einer Dimension an, die kleine MSE erzielt hat, und vergleicht diese mit dem Restplot der Neurotizität, sieht man, dass die vorhergesagten Werte in einem viel kleineren Bereich sind. Die Reste sind ebenfalls kleiner.

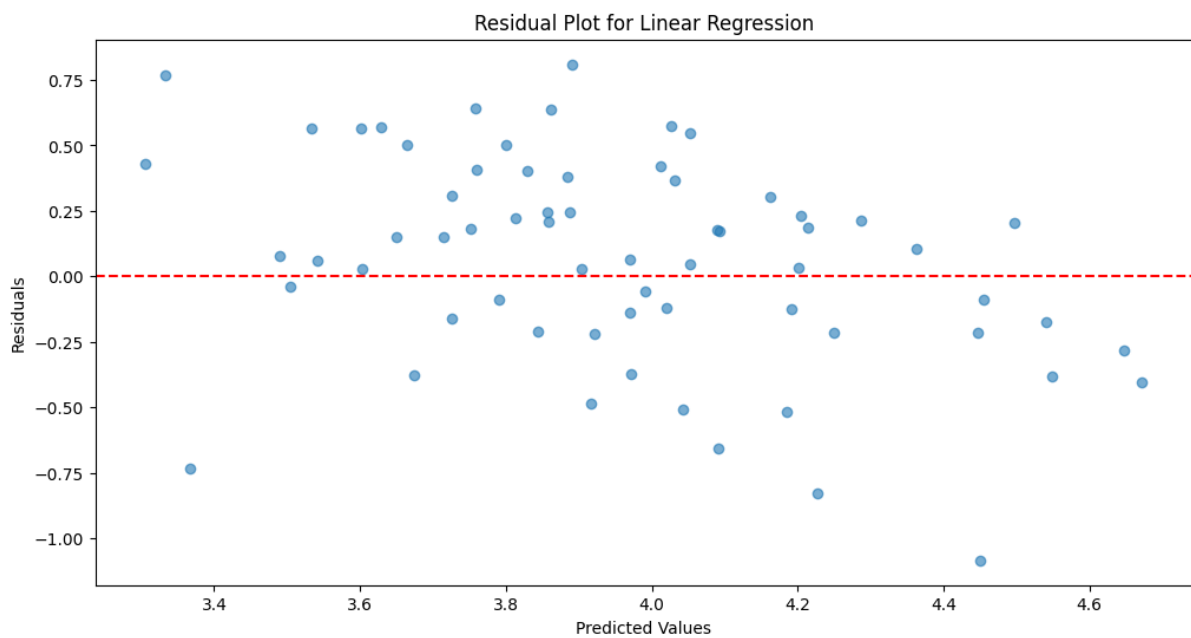


Abbildung 29: Restplot der Ergebnisse der Vorhersage der Kategorie Verträglichkeit

## 7. Diskussion und Ausblick

Grundsätzlich kann man sagen, dass das Ziel, BFAS-Daten anhand von Ressourcenbilderdaten vorauszusagen nicht erreicht wurde. Weder im Clustering, noch bei den Prediction Models konnte ein signifikanter Zusammenhang zwischen den Ressourcenbildern und den BFAS Dimensionen nachgewiesen werden.

Wenn man das Ganze aus psychologischer Sicht betrachtet, stellt sich die Frage, inwiefern es ein Unterschied ist, ob beispielsweise der Neurotizismus einen Score von 3.8 oder 4 ist. Wenn man so argumentiert, kann die BFAS-Skala beispielsweise in tiefe und hohe Werte geteilt werden. Dann könnte eine Kategorisierung mittels logistischer Regression versucht werden, anstatt numerische Werte mit linearer Regression vorherzusagen. Auch Random Forest Modelle, welche im Rahmen dieser Arbeit lineare Regression bereits übertroffen haben, wären hierfür gut geeignet. Weiterhin ist fraglich, ob die Ressourcenbilder für diesen Versuch geeignet sind. In der Literatur zu den Ressourcenbilder steht, dass sie ein Werkzeug seien, um daraus relevante Lebensthemen, Bedürfnisse und Motive zu identifizieren. [19] Ob diese Lebensthemen einen Einfluss auf die Big Five Dimensionen haben, stellt sich hier in Frage. In den beschriebenen Studien im Kapitel 3.2 hatten die Inputdaten gefühlswise mehr Relevanz in Bezug auf die Big Five Dimensionen, als die Ressourcenbilder.

Ein weiterer Punkt sind die Daten selbst. Die vorhandene Datenmenge war mit 324 Samples klein. Die Studie von Dominik Reuter [16] welche die Smartphone Daten analysiert hat, hatte 1523 Teilnehmer [16] und die Studie, welche die Portraits analysiert hat, 30935 [20]. Die verschiedenen Versionierungen der Umfragen haben dabei sicher auch nicht geholfen, es war zum Teil schwierig zu akzeptieren, die umformulierten Fragen als identisch zu behandeln, so wie sie verändert wurden.

Wenn man eine etwas grössere Studie anschaut, bei der mehr als eine Million Daten gesammelt wurden, sieht man die Verteilung der Big Five Scores in grossen Populationen. [21]

**Table 4.** Average OPQ32 Big Five Sten Scores for the 31 Countries: Means and SDs (for Sample Sizes, See Table 3), Together With the Average SD Across the Five Scales (Final Column)

Country	Language	Emotional Stability		Extraversion		Openness		Agreeableness		Conscientiousness		M SD
		M	SD	M	SD	M	SD	M	SD	M	SD	
Argentina	Spanish	5.47	1.59	5.98	1.66	5.21	1.74	4.81	1.77	5.51	1.74	1.70
Australia	English	5.58	2.03	5.39	2.06	5.40	2.01	5.68	1.95	5.68	2.00	2.01
Belgium	French	4.87	2.00	5.53	1.89	5.59	1.91	5.67	1.88	5.73	2.06	1.95
Belgium	Flemish	5.42	2.18	5.86	2.14	5.48	2.05	6.21	1.94	5.38	2.08	2.08
Brazil	Portuguese	4.91	1.69	5.63	1.86	5.12	1.88	5.58	1.73	4.75	1.80	1.79
Canada	French	5.83	1.86	6.39	1.89	5.24	1.92	5.68	1.87	5.64	1.92	1.89
China (PRC)	Simplified Chinese	5.46	1.91	5.13	1.91	5.03	1.79	5.91	1.61	5.36	1.75	1.79
Denmark	Danish	6.38	1.82	6.37	1.98	5.60	2.01	5.76	1.86	5.51	2.06	1.95
Finland	Finnish	6.22	2.06	5.94	2.24	5.01	2.11	6.08	1.98	5.44	2.14	2.11
France	French	4.85	2.00	5.47	1.84	5.67	1.92	5.62	1.76	5.55	2.00	1.90
Germany	German	6.08	2.00	5.96	1.88	6.08	1.86	5.87	1.80	5.32	1.84	1.88
Greece	Greek	4.32	1.90	5.76	1.67	5.54	1.97	4.97	1.80	5.87	2.10	1.89
Hong Kong	Traditional Chinese	5.13	2.16	5.43	2.09	5.58	1.95	5.67	1.87	5.36	1.84	1.98
Hungary	Hungarian	5.73	2.08	5.21	1.96	4.97	2.07	5.52	1.81	5.43	2.12	2.01
India	English	5.41	1.72	5.18	1.74	5.16	1.79	4.19	1.69	5.52	1.74	1.74
Indonesia	Indonesian	4.98	1.67	5.02	1.81	5.08	1.75	5.18	1.67	4.93	1.86	1.75
Italy	Italian	5.09	2.00	5.70	1.85	5.86	1.93	5.00	1.82	5.02	1.89	1.90
Japan	Japanese	4.46	2.27	5.16	1.91	6.45	1.84	5.62	1.78	4.15	1.86	1.93
Malaysia	English	4.90	1.81	4.58	1.92	5.11	1.74	5.29	1.67	5.17	1.93	1.82
Middle East	Arabic	4.98	1.51	4.86	1.44	5.33	1.68	4.96	1.62	5.18	1.69	1.59
Netherlands	Dutch	6.01	2.05	5.84	1.98	6.08	1.93	5.73	1.84	4.49	1.89	1.94
New Zealand	English	5.61	1.98	5.20	1.95	5.27	1.98	5.91	1.88	6.18	2.06	1.97
Norway	Norwegian	6.22	1.83	6.29	1.91	5.30	1.88	6.14	1.91	5.26	1.89	1.88
Poland	Polish	4.99	2.01	4.64	2.03	5.84	1.81	4.75	1.69	5.35	2.07	1.92
Portugal	Portuguese	4.77	1.76	5.88	1.79	5.79	2.01	5.81	1.68	5.12	1.96	1.84
Russia	Russian	4.76	1.96	5.27	1.76	6.19	1.98	5.10	1.78	5.02	1.86	1.87
South Africa	English	5.28	1.89	4.93	1.94	5.45	1.89	5.06	1.82	5.60	1.97	1.90
Spain	Spanish	4.79	1.78	5.75	1.90	5.61	1.88	5.40	1.80	5.13	1.94	1.86
Sweden	Swedish	6.24	1.90	6.37	1.91	5.61	1.87	6.28	1.89	5.83	2.00	1.91
United Kingdom	English	5.38	2.04	5.52	2.00	5.66	2.03	5.79	1.94	5.61	2.05	2.01
United States	English	5.70	2.05	5.84	2.09	5.29	2.05	5.34	1.97	5.72	2.03	2.04

Abbildung 30: Durchschnittliche Big Five Verteilung für 31 Ländern mit Mittelwert und Standardabweichung, Skala 1-7 [21]

	Mittelwert	Median	Varianz
Extraversion	3.57	3.60	0.26
Neurotizismus	2.69	2.67	0.46
Offenheit/Intellekt	3.77	3.77	0.22
Gewissenhaftigkeit	3.53	3.56	0.26
Verträglichkeit	3.96	4.00	0.19

Tabelle x: BFAS kombiniert Mittelwerte, Median und Varianz. Skala 1-5

Wenn man sich die obere Tabelle anschaut, sieht man, dass es durchaus Unterschiede in den verschiedenen Ländern gibt. Vergleicht man die Verteilung aus Deutschland beispielsweise mit den uns verfügbaren BFAS Werte, so sieht man, dass es doch wesentliche Unterschiede gibt. In der Neurotizität sieht man zum Beispiel einen Unterschied von 2.7 (auf einer Skala von 1 bis 5)

und 6 (auf einer Skala von 1 bis 7). Auch die hohe Verträglichkeit ist auffällig. Dies kann ein Zeichen sein, dass zu wenig Daten zur Verfügung standen.

### Ausblick

Während der Arbeit kamen weitere Ideen für eine zukünftige Folgearbeit aus. Vermutlich sollten davor aber noch mehr einheitliche Daten zur Verfügung stehen. Folgende Ideen kamen auf:

- Andere Modelle verwenden
- Mit einer Klassifizierung versuchen

## 8. Literaturverzeichnis

- [1] M. Schreiber, R. Mäder and M. Inniger. [Online]. Available: <https://www.laufbahndiagnostik.ch/fragebogen/Laufbahn>. [Accessed 06 06 2024].
- [2] Institut für Angewandte Psychologie, "Arbeitsmittel," [Online]. Available: <https://www.laufbahndiagnostik.ch/tools>. [Accessed 06 06 2024].
- [3] Institut für Angewandte Psychologie, "Laufbahndiagnostik," [Online]. Available: <https://www.laufbahndiagnostik.ch/de>. [Accessed 06 06 2024].
- [4] M. Schreiber and G. J. Jenny, "Development and validation of the 'Lebender emoticon PANAVA' scale (LE-PANAVA) for digitally measuring positive and negative activation, and valence via emoticons," *Personality and Individual Differences*, vol. 160, 2020.
- [5] H. A. Murray, Thematic apperception test manual., Harvard University Press., 1943.
- [6] L. R. Goldberg, "An alternative «description of personality»: The Big-Five factor," *Journal of Personality and Social Psychology*, vol. 59, p. 1216–1229, 1990.
- [7] C. G. DeYoung, L. Quilty and J. B. Peterson, "Between Facets and Domains: 10 Aspects of the Big Five," *Journal of Personality and Social Psychology*, vol. 93, no. 5, p. 880–896, 2007.
- [8] P. Mussel and M. Paelecke, "BFAS-G. Big Five Aspect Scales - German," ZPID (Leibniz Institute for Psychology Information) – Testarchiv, 2018. [Online]. Available: <https://www.testarchiv.eu/de/test/9007737>. [Accessed 1 6 2024].
- [9] M. I. M. S. Marie-Louise Iller, "Laufbahndiagnostik," [Online]. Available: [https://www.laufbahndiagnostik.ch/downloads/de/Handbuch\\_BFAS-G-R\\_2023-09.pdf](https://www.laufbahndiagnostik.ch/downloads/de/Handbuch_BFAS-G-R_2023-09.pdf). [Accessed 01 06 2024].
- [10] C. G. DeYoung, "Cybernetic Big Five Theory," *Journal of Research in Personality*, vol. 56, pp. 33-58, 2015.

- [11] M. Schreiber, G. J. Jenny, M. Hürlimann, Y. Parfenova, P. v. Däniken and M. C. Paper], “Exploring the use of artificial intelligence (AI) in personality psychology: Can AI connect idiographic narratives with nomothetic questionnaire data?,” no. [Unreleased].
- [12] S. K. a. T. v. O. Karl-Heinz Renner, “Bringing Back the Person into Behavioural Personality Science Using Big Data,” *European Journal of Personality*, vol. 34, pp. 670 - 686, 13 09 2020.
- [13] J. F. R. Le Vy Phan, “Personality computing: New frontiers in personality assessment,” 11 05 2021.
- [14] “Medi-Karriere,” 25 05 2024. [Online]. Available: <https://www.medi-karriere.ch/magazin/rorschach-test-methodik-anwendung-und-analyse/>.
- [15] M. S. a. M. T. Bruce Ferwerda, “Using Instagram Picture Features to Predict,” *MultiMedia Modeling*, pp. 850-861, 2016.
- [16] M. S. M. N. M. A. E. F. u. T. K. Dominik Rügger, “How Are Personality States Associated with Smartphone Data?,” *SageJournals*, 01 09 2020.
- [17] A. Sattiraju, S. Roy and D. Viji, “Personality Prediction in Candidates using a Picture Based Test,” in *Artificial Intelligence Techniques for Advanced Computing Applications. Lecture Notes in Networks and Systems, vol 130. Springer, Singapore.*, Singapore, Springer, 2020, pp. 445-455.
- [18] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.
- [19] M. Schreiber, *Narrative Ansätze in Beratung und Coaching*, Springer.
- [20] M. C. M.-S. Moreno-Armendáriz, “Estimation of Personality Traits From Portrait,” 25 05 2024. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9244051>.
- [21] D. Bartram, *Scalar Equivalence of OPQ32: Big Five Profiles of 31 Countries*, SAGE Publications, 2013.

## 9. Abbildungsverzeichnis

Abbildung 1: Screenshot LE-PANA-M, Erfassen der Gemütslage .....	6
Abbildung 2: Beispiel für ein Ressourcenbild .....	7
Abbildung 3: BFAS-G-R Frageschema .....	9
Abbildung 4: Ergebnisse der klassischen Textanalyseverfahren mit Supervised Prediction Models auf BFAS Dimensionen .....	11
Abbildung 5: Ausschnitt aus «BFAS-G-R_Items zwei Versionen.xlsx». Zu sehen sind drei Items, von denen zwei angepasst wurden, siehe Spalte «Change». Bei einem Item wurde ausserdem die Polung umgekehrt, siehe Spalte «Polung r = reversed». ....	16

Abbildung 6: Sichtbare Unterschiede zwischen BFAS-G und BFAS-G-R aufgrund fehlerhafter Umpolung.....	16
Abbildung 7: Zwei Einträge im Datensatz, welche zu verschiedenen Versionen gehören. ....	18
Abbildung 8: Histogramm der Verteilung der Dimension Neurotizismus. ....	21
Abbildung 9: .....	21
Abbildung 10: Boxplots der fünf BFAS-Dimensionen .....	21
Abbildung 11: Absolute Differenz der Durchschnittswerte zwischen BFAS und BFAS-G-R nach Itemstatus. ....	22
Abbildung 12: Geschlechtliche Unterschiede der BFAS-Aspekte.....	23
Abbildung 13: Abweichung der Durchschnittlichen BFAS-Aspekte der persönlichen Daten von den Basiswerten .....	24
Abbildung 14: Beliebteste Ressourcenbilder, nach absteigender Beliebtheit sortiert.....	25
Abbildung 15: Unbeliebteste Ressourcenbilder, nach aufsteigender Beliebtheit sortiert. ....	25
Abbildung 16: Spärliche Werteverteilung der Kategorie relax.....	25
Abbildung 17: Geschlechtliche Unterschiede der Ressourcenbilderpräferenzen. ....	26
Abbildung 18: Dimensional reduzierte BFAS-Daten, nach Version gefärbt.....	30
Abbildung 19: Verlauf der Silhouette Scores in Abhängigkeit von k. ....	31
Abbildung 20: Silhouette Plot der Ressourcenbilder .....	32
Abbildung 21: Contourplot der DBSCAN optimierung auf zweidimensionaler PCA-Projektion der Aspekte .....	33
Abbildung 22: Testcase 1, k =5 Crossvalidation, MSE .....	35
Abbildung 23: Testcase 2, k =5 Crossvalidation, MSE .....	36
Abbildung 24: Testcase 3, k =5 Crossvalidation, MSE .....	37
Abbildung 25: Testcase 1, Leave-One-Out Crossvalidation, MSE.....	37
Abbildung 26: Testcase 1, Leave-One-Out Crossvalidation, PCA, MSE.....	38
Abbildung 27: Verteilung von der MSE von 100 verschiedenen Durchläufen der Vorhersagen... ..	39
Abbildung 28: Restplot der Ergebnisse der Vorhersage der Kategorie Neurotizismus .....	40
Abbildung 29: Restplot der Ergebnisse der Vorhersage der Kategorie Verträglichkeit .....	40
Abbildung 30: Durchschnittliche Big Five Verteilung für 31 Ländern mit Mittelwert und Standardabweichung, Skala 1-7 [21] .....	42
Abbildung 31:.....	46
Abbildung 32: Durchschnittliche BFAS Werte nach Dimension, BFAS-G vs BFAS-G-R .....	46
Abbildung 33: Durchschnittliche BFAS Werte nach Aspekt, BFAS-G vs BFAS-G-R .....	47
Abbildung 34: Testcase 1, k=5 Crossvalidation, MSE.....	47
Abbildung 35: Testcase 1, Leave-One-Out Crossvalidation, MSE.....	48
Abbildung 36: Testcase 1, k=5 Crossvalidation, PCA, MSE.....	48
Abbildung 37: Testcase 1, Leave-One-Out Crossvalidation, PCA, MSE.....	49
Abbildung 38: Testcase 2, k=5 Crossvalidation, MSE.....	49
Abbildung 39: Testcase 2, Leave-One-Out Crossvalidation, MSE.....	50
Abbildung 40: Testcase 2, k=5 Crossvalidation, PCA, MSE.....	50
Abbildung 41: Testcase 2, Leave-One-Out Crossvalidation, PCA, MSE.....	51
Abbildung 42: Testcase 3, k=5 Crossvalidation, MSE .....	52
Abbildung 43: Testcase 3, Leave-One-Out Crossvalidation, MSE .....	52
Abbildung 44: Testcase 3, k=5 Crossvalidation, PCA, MSE.....	53
Abbildung 45: Testcase 3, Leave-One-Out Crossvalidation, PCA, MSE.....	53
Abbildung 46: Testcase 4, k=5 Crossvalidation , MSE .....	54
Abbildung 47: Testcase 4, Leave-One-Out Crossvalidation , MSE .....	54
Abbildung 48: Testcase 4, k=5 Crossvalidation, PCA, MSE.....	55

## 10. Technische Dokumentation

Die Arbeit wurde hauptsächlich in Python verfasst, genauer gesagt in Jupyter Notebooks. Die verschiedenen Notebooks sind unterteilt in die Prozessschritte und dementsprechend benannt. Es werden nur Python Librarys benötigt, die normal mit *pip install* installiert werden können. Die Daten sind ebenfalls in Form von csv Dateien im Projekt Repo enthalten. Hinweis zu XGBoost: Mit `XGBRegressor(device="cuda")` kann die Grafikkarte ebenfalls genutzt werden.

## 11. Anhang

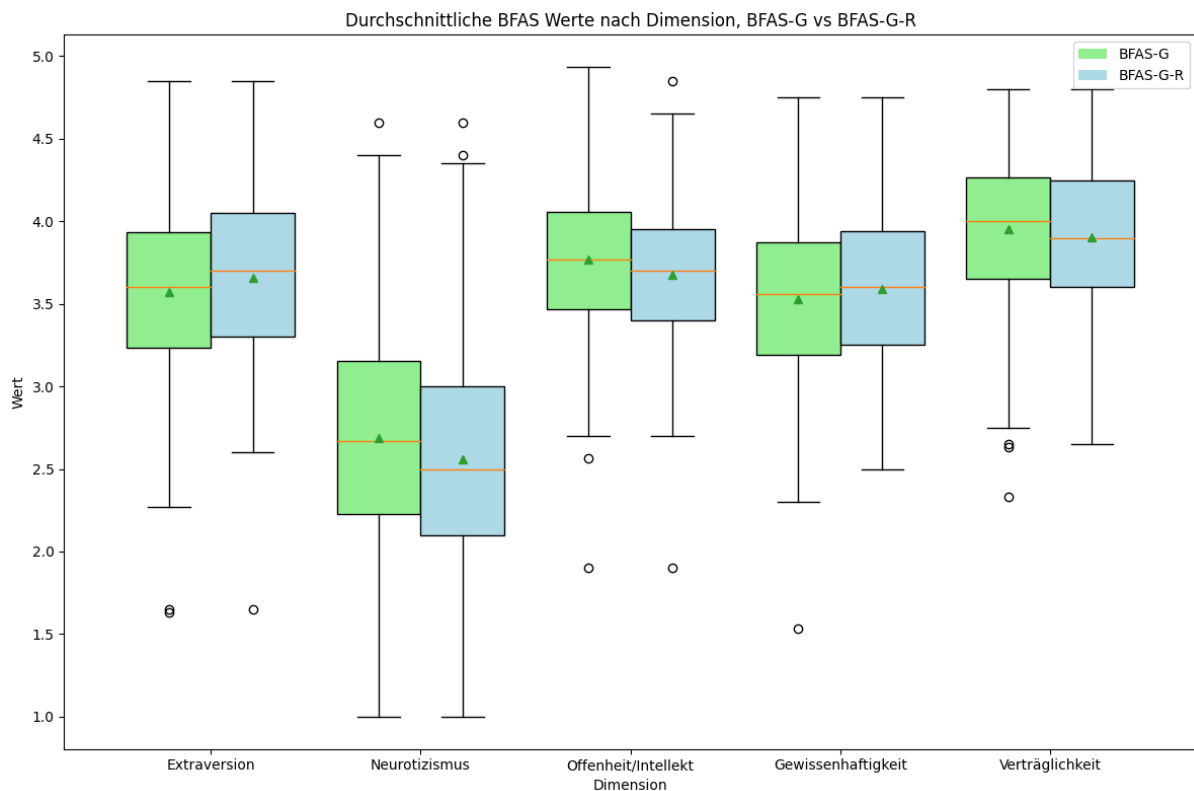


Abbildung 32: Durchschnittliche BFAS Werte nach Dimension, BFAS-G vs BFAS-G-R

Abbildung 31:

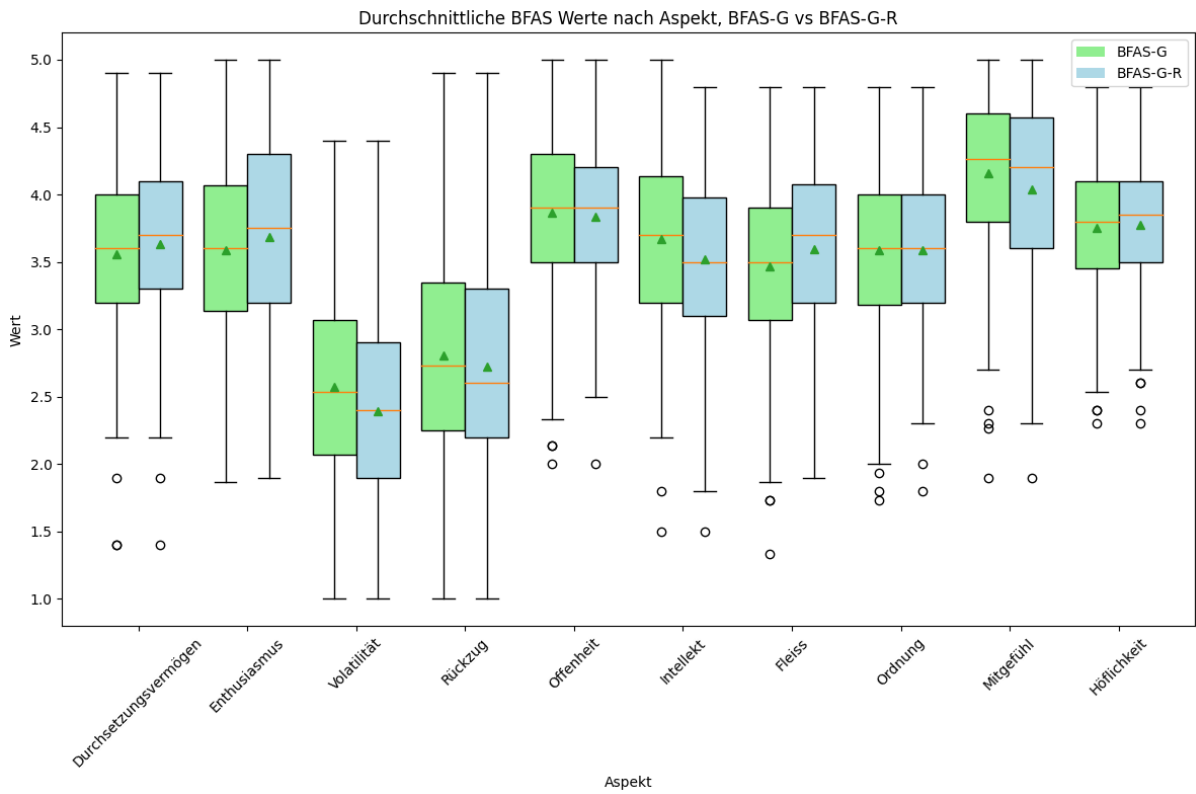


Abbildung 33: Durchschnittliche BFAS Werte nach Aspekt, BFAS-G vs BFAS-G-R

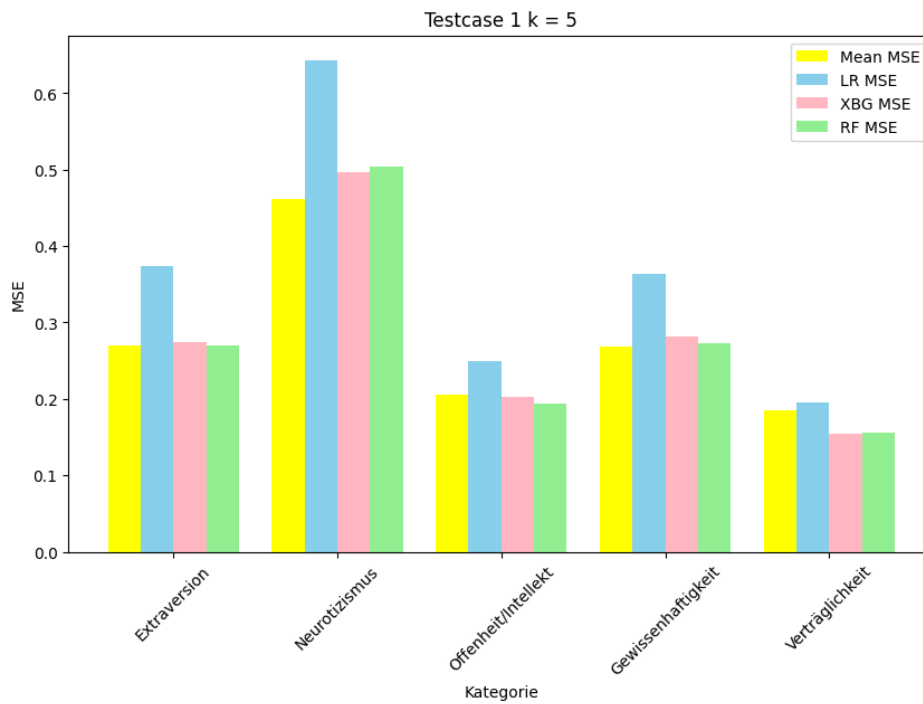


Abbildung 34: Testcase 1, k=5 Crossvalidation, MSE



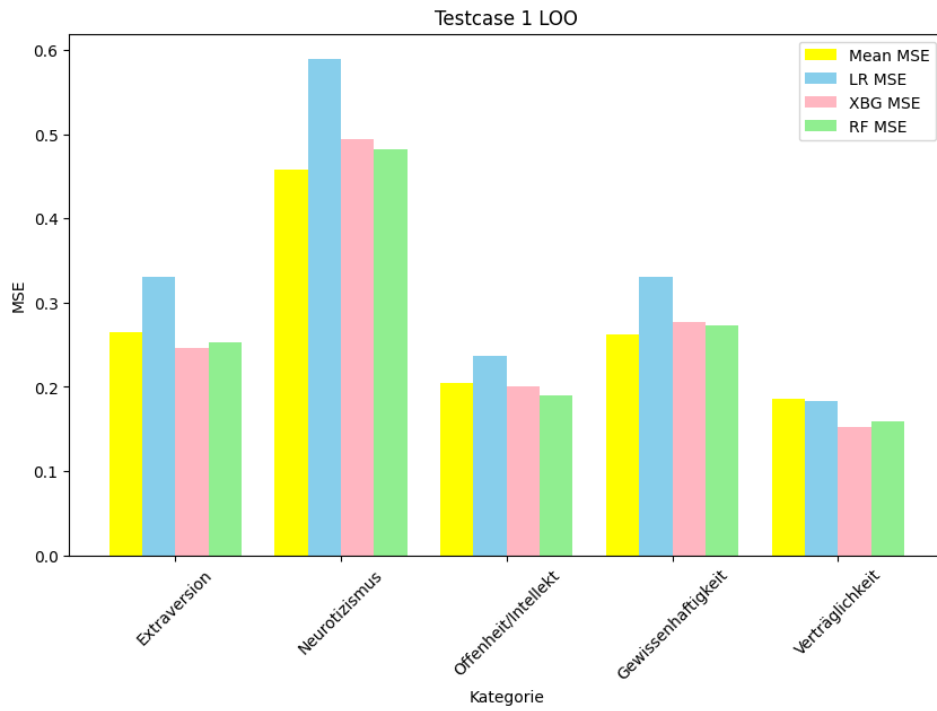


Abbildung 35: Testcase 1, Leave-One-Out Crossvalidation, MSE

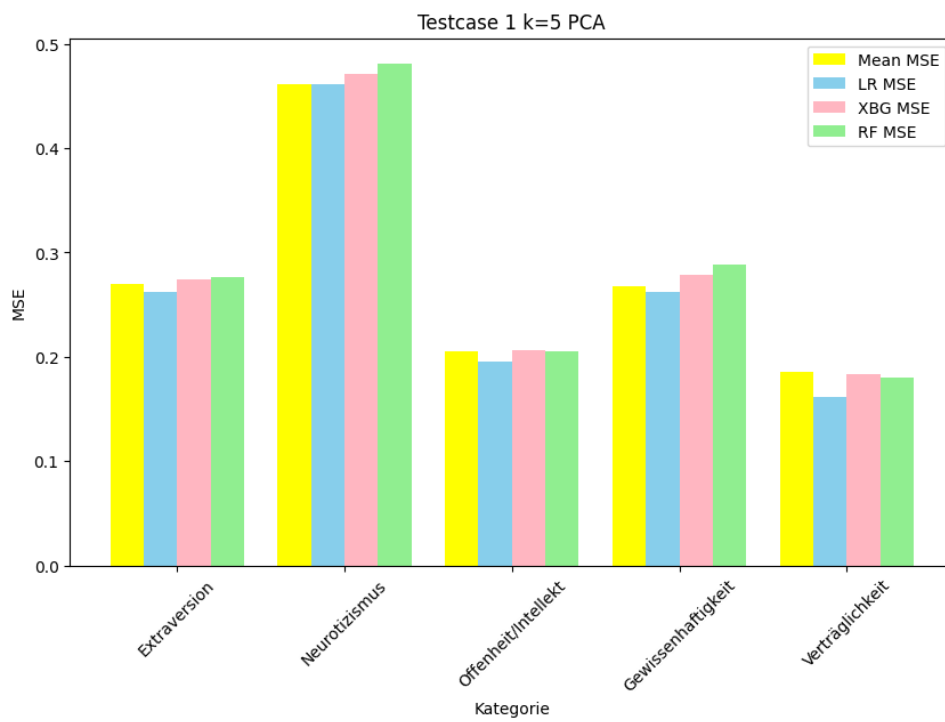


Abbildung 36: Testcase 1, k=5 Crossvalidation, PCA, MSE

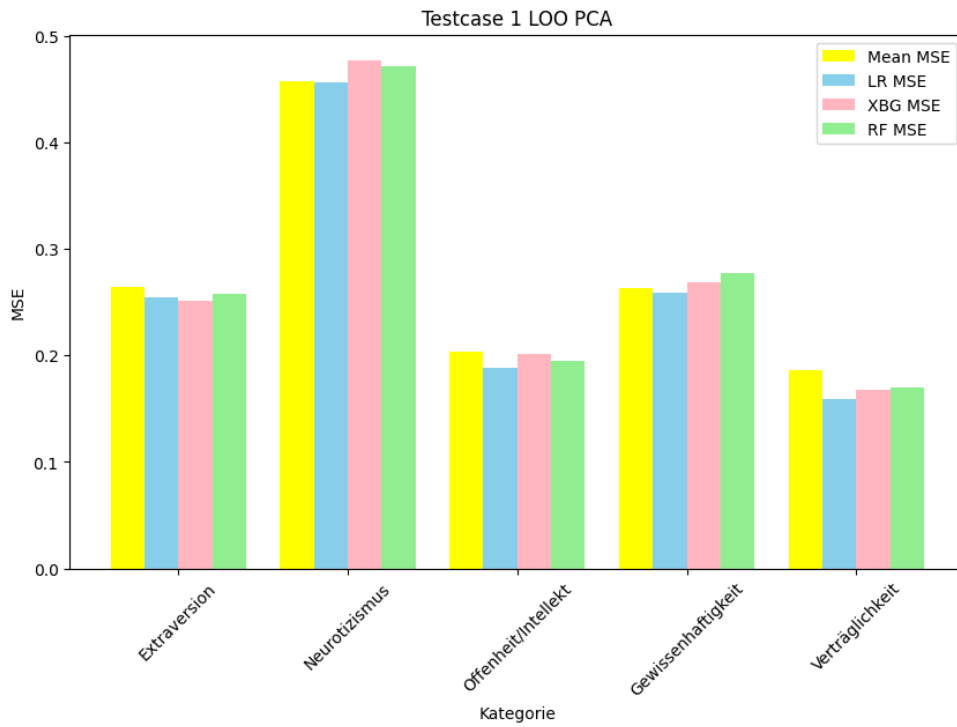


Abbildung 37: Testcase 1, Leave-One-Out Crossvalidation, PCA, MSE

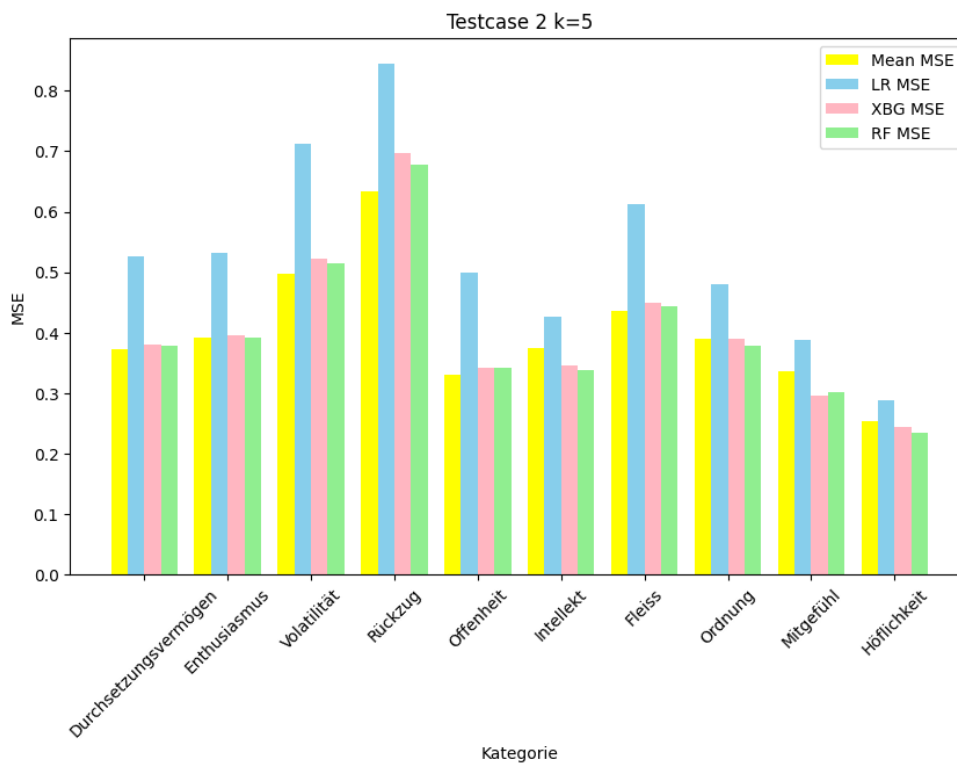


Abbildung 38: Testcase 2, k=5 Crossvalidation, MSE

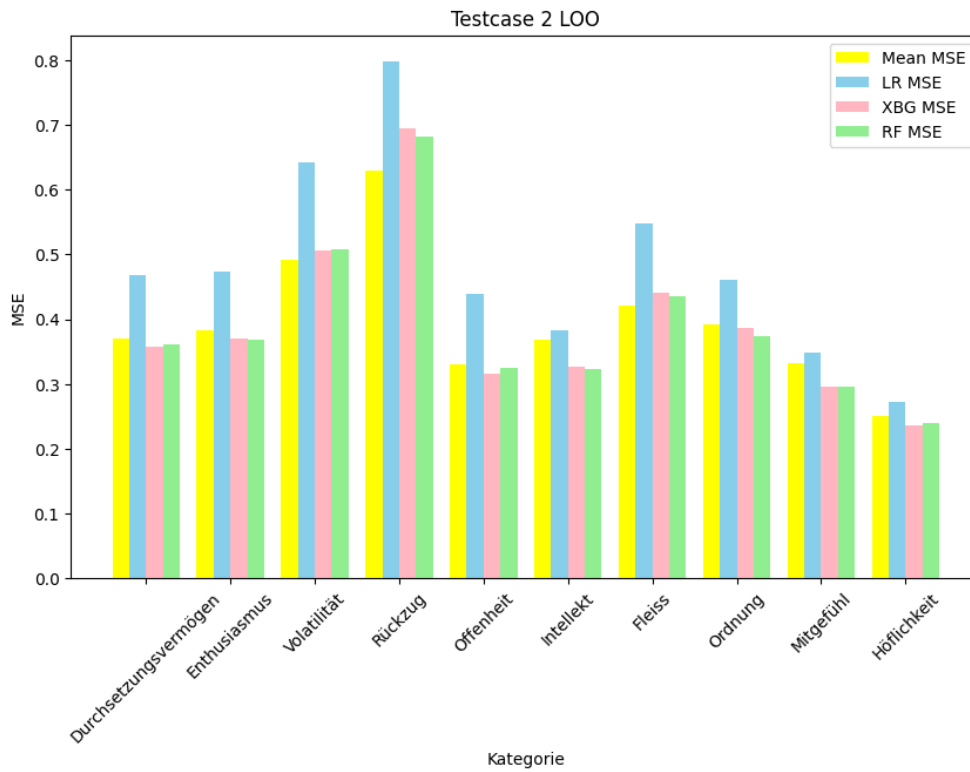


Abbildung 39: Testcase 2, Leave-One-Out Crossvalidation, MSE

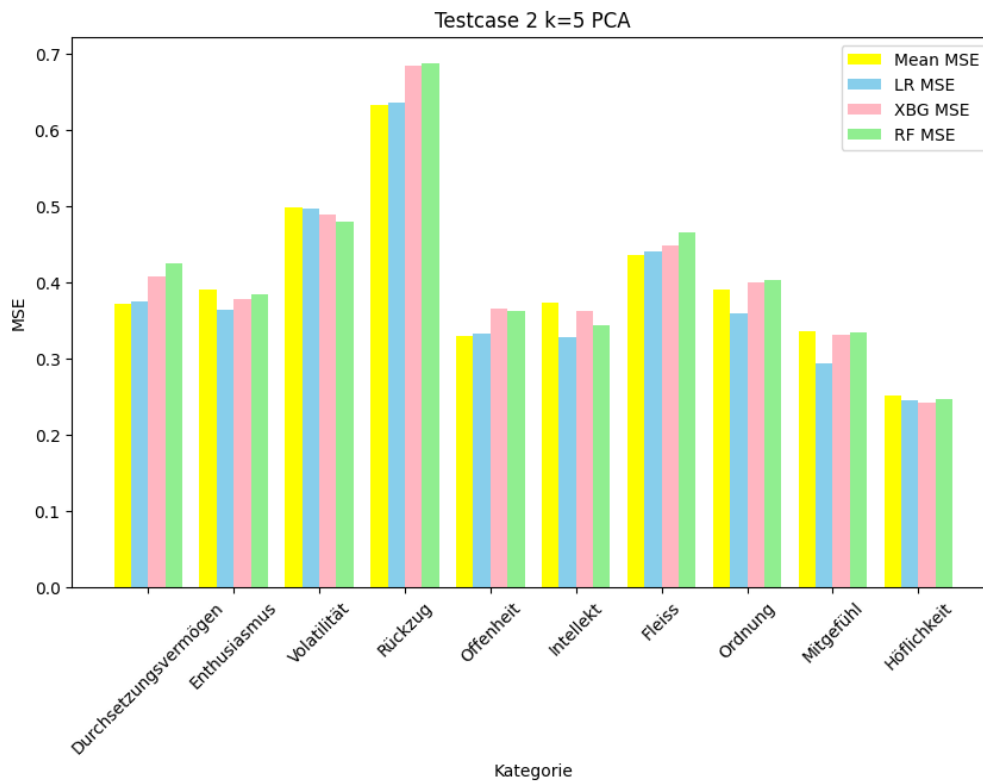


Abbildung 40: Testcase 2, k=5 Crossvalidation, PCA, MSE

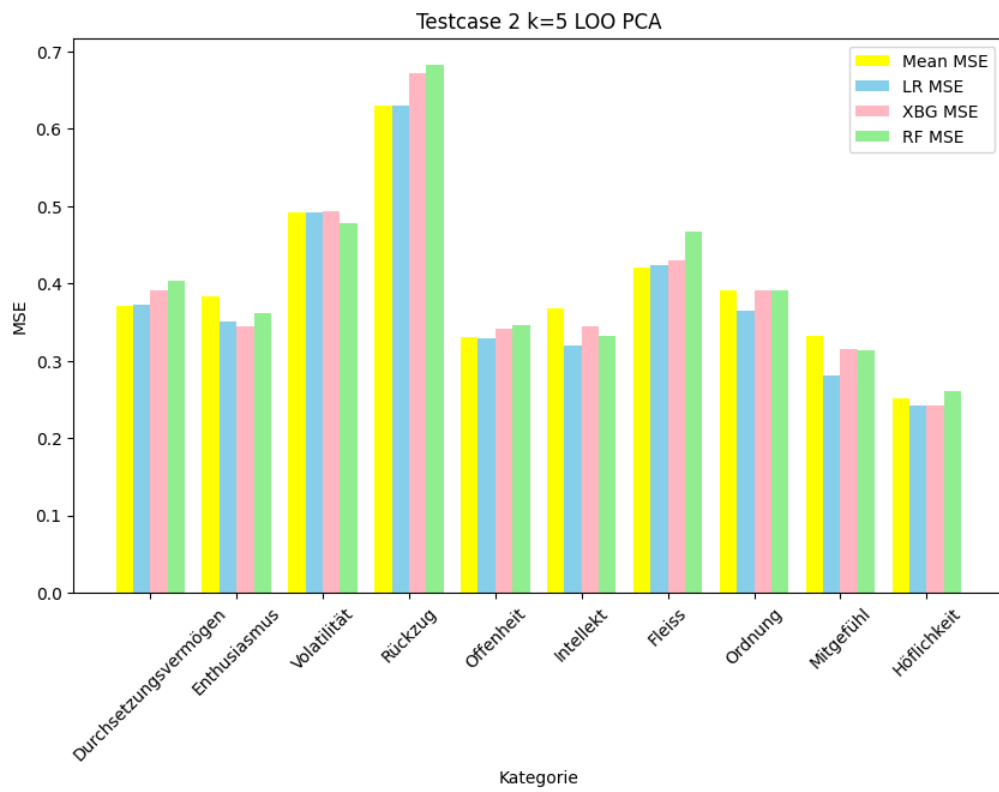


Abbildung 41: Testcase 2, Leave-One-Out Crossvalidation, PCA, MSE

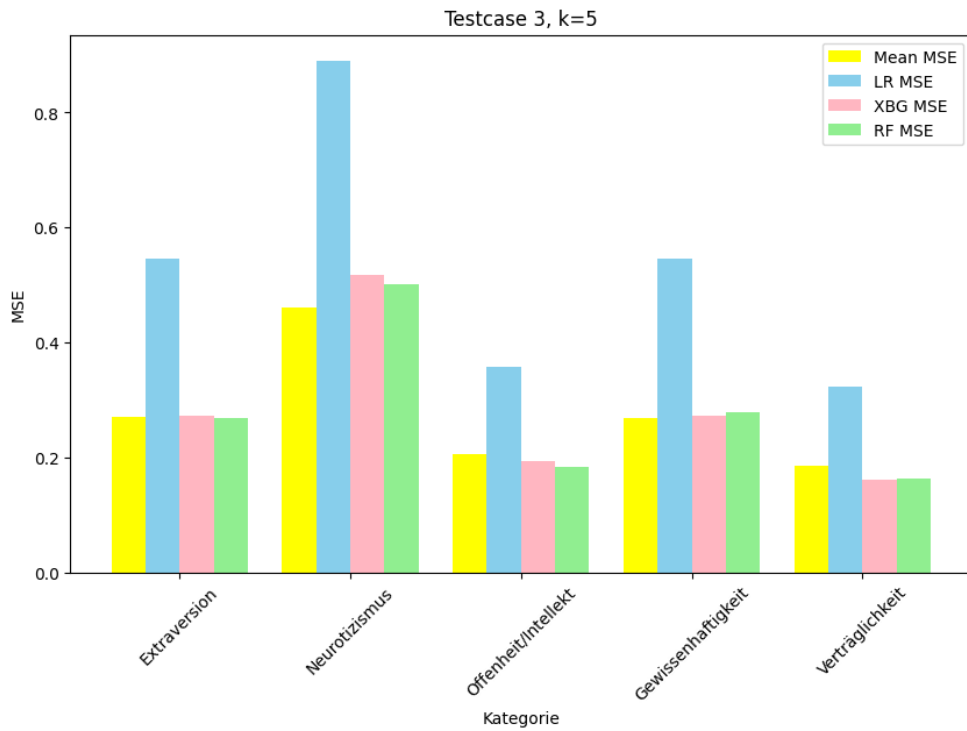


Abbildung 42: Testcase 3, k=5 Crossvalidation, MSE

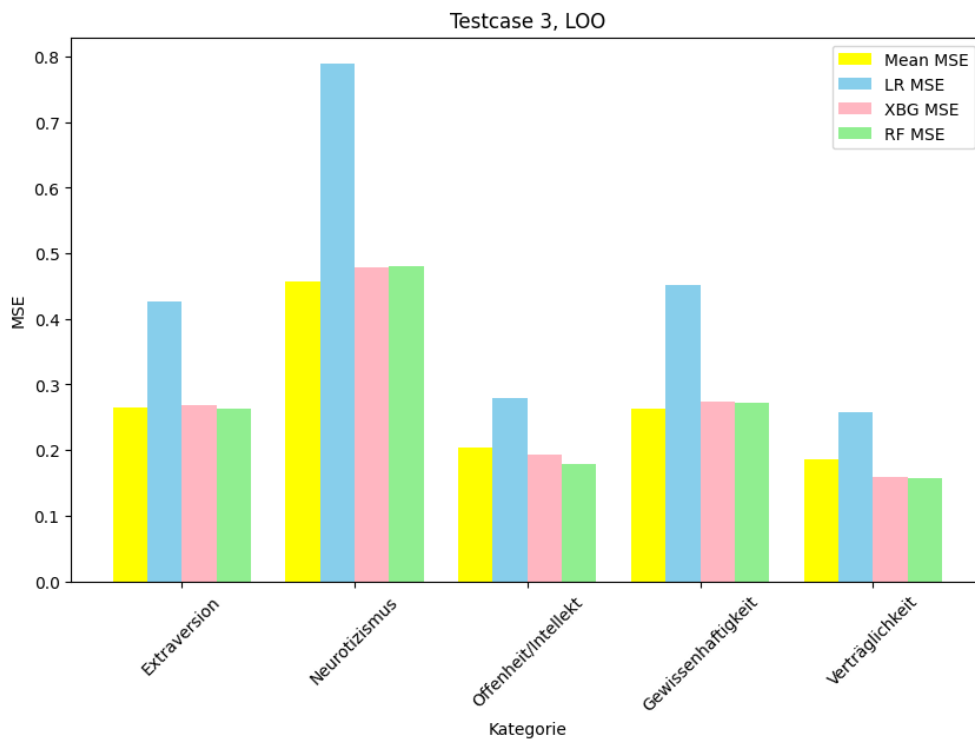


Abbildung 43: Testcase 3, Leave-One-Out Crossvalidation, MSE

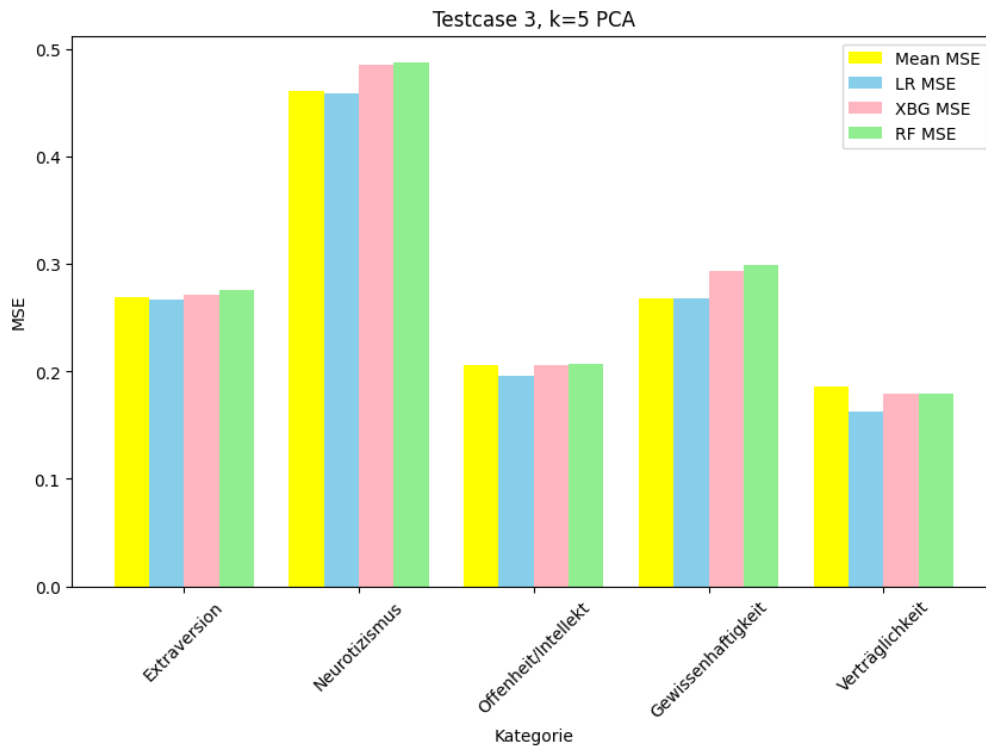


Abbildung 44: Testcase 3, k=5 Crossvalidation, PCA, MSE

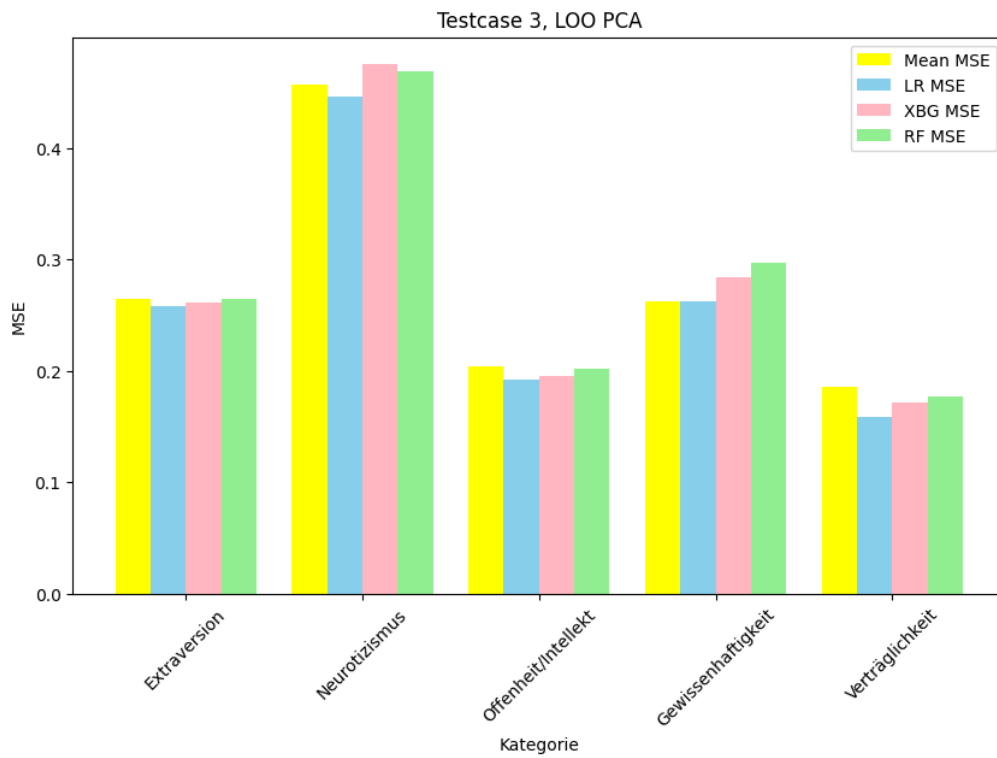


Abbildung 45: Testcase 3, Leave-One-Out Crossvalidation, PCA, MSE

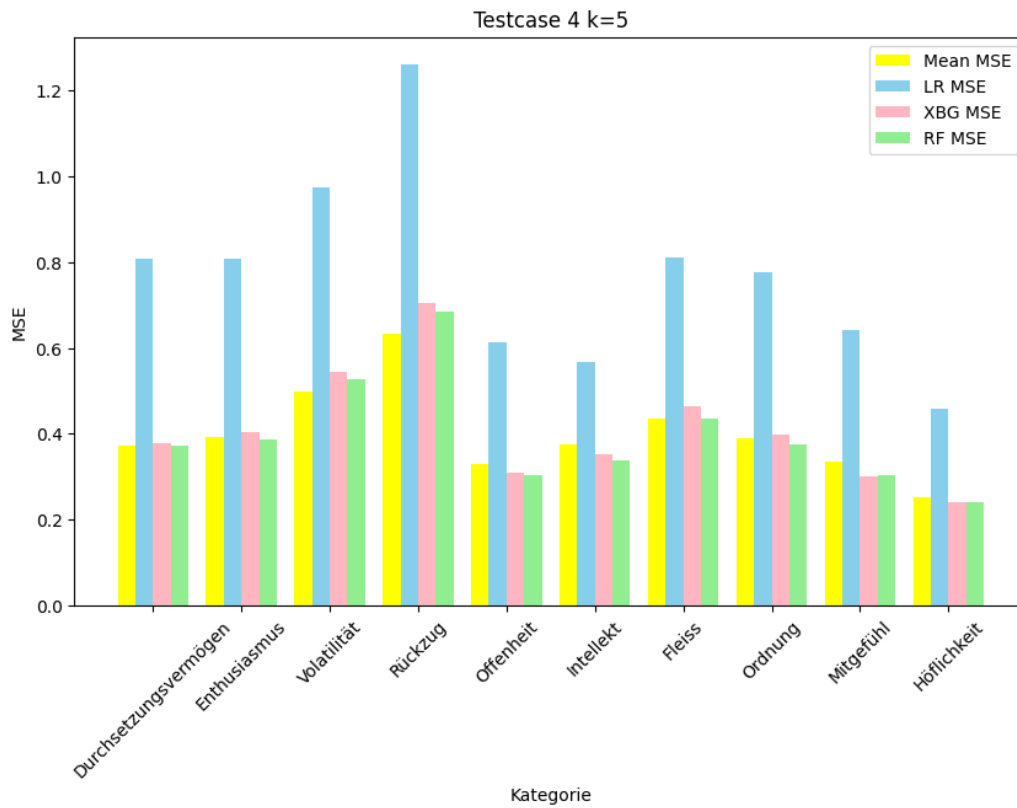


Abbildung 46: Testcase 4, k=5 Crossvalidation , MSE

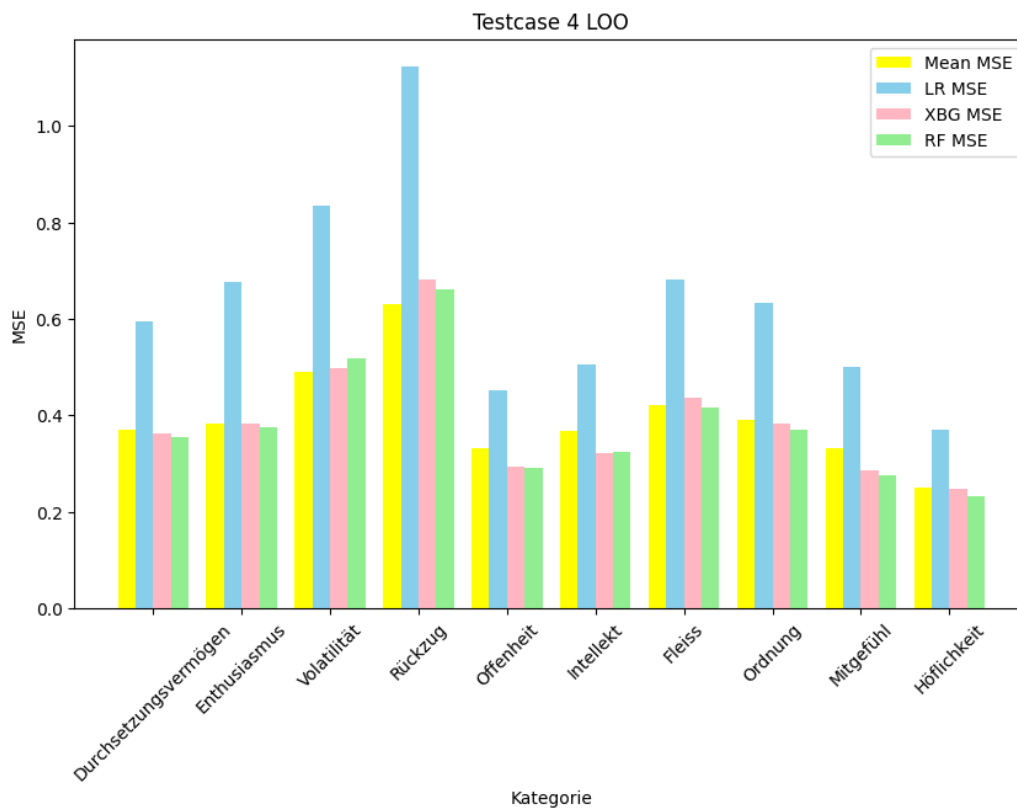


Abbildung 47: Testcase 4, Leave-One-Out Crossvalidation , MSE

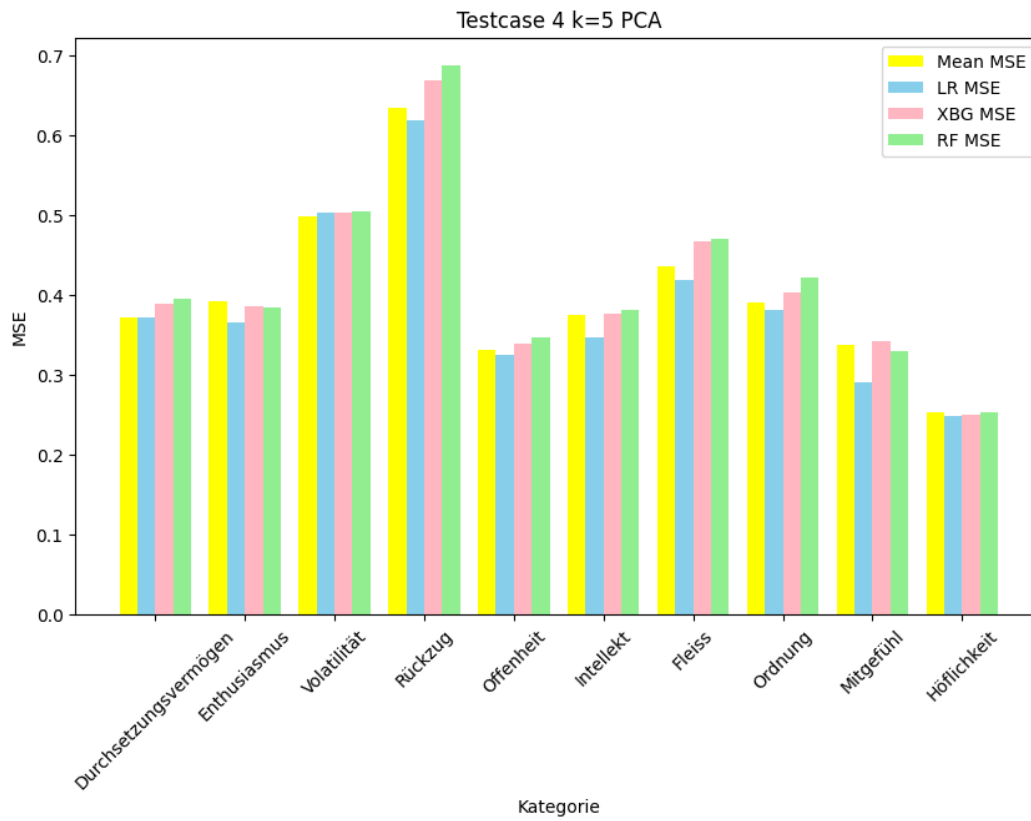


Abbildung 48: Testcase 4, k=5 Crossvalidation, PCA, MSE

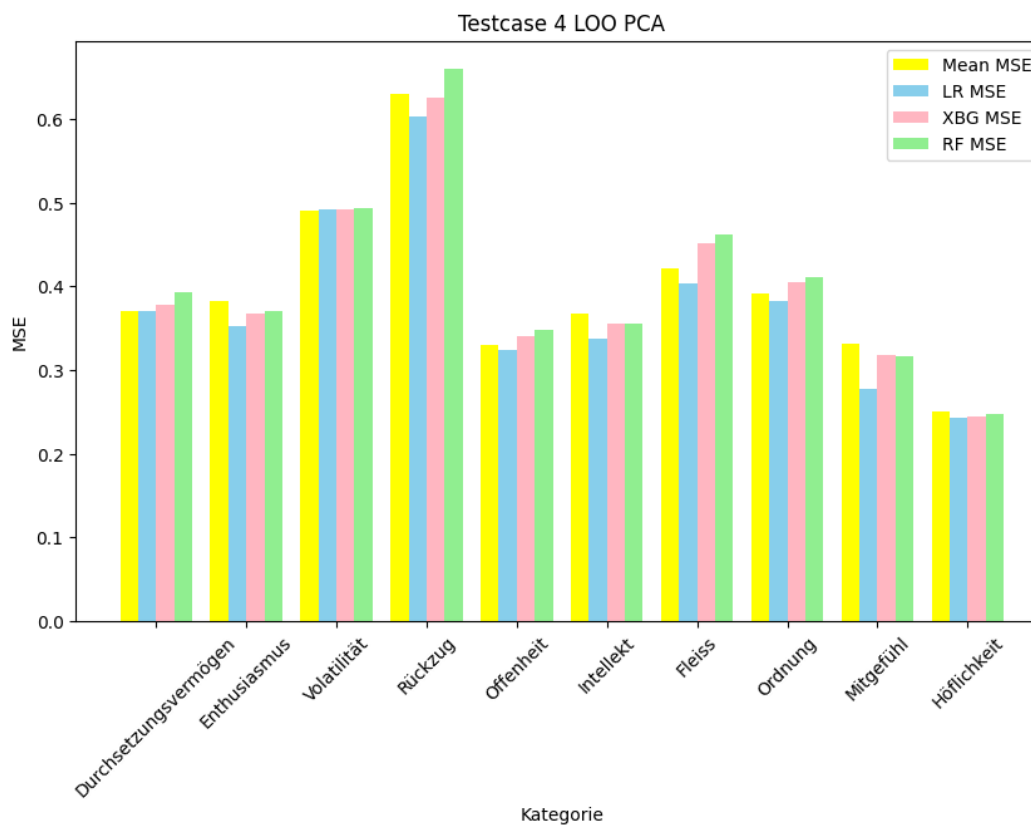


Abbildung 49: Testcase 4, Leave-One-Out Crossvalidation, PCA, MSE