Fallstudie

Psychische Gesundheit in technologiebezogenen Berufen

Studiengang
Ängewandte Künstliche Intelligenz”

Maschinelles Lernen - Unsupervised Learning und Feature Engineering
DLBDSMLUSL01_D

Dawid Jedlinski

Matrikelnummer: IU14113900

dawid.jedlinski@iu-study.org

Tutor: Christian Müller-Kett

Abgabedarum: dd.mm.yyyy

Inhaltsverzeichnis

I. Abbildungsverzeichnis

II. Tabellenverzeichnis

III. Abbreviations

MH	Mental Health (Mentale Gesundheit)
OSMI	Open Sourcing Mental Illness
EDA	Explorative Datenanalyse
PCA	Principal Component Analysis
MDS	Multidimensional Scaling
LLE	Locally Linear Embedding
GMM	Gaussian Mixture Model
HR	Human Resources
NLP	Natural Language Processing
ML	Machine Learning
CPU	Central Processing Unit
CRED	C Range Error Detector
Dev	Development, the development team

1. Einleitung

Die fortschreitende Digitalisierung sowie die zunehmende Verdichtung von Arbeitsprozessen stellen Unternehmen vor neue Herausforderungen im Bereich des betrieblichen Gesundheitsmanagements. Insbesondere in technologieintensiven Berufsfeldern, die durch hohe kognitive Anforderungen und komplexe Problemlösungsaufgaben geprägt sind, gewinnt die psychische Gesundheit der Beschäftigten zunehmend an Bedeutung. Vor diesem Hintergrund implementieren Organisationen verstärkt interne Präventionsprogramme. Deren Erfolg hängt jedoch entscheidend davon ab, wie die erhobenen Daten interpretiert und in konkrete, zielgerichtete Maßnahmen überführt werden können. Die vorliegende Arbeit beschäftigt sich mit der analytischen Aufbereitung und Auswertung solcher Daten unter Anwendung fortgeschrittener Methoden des Data Science.

1.1. Problemstellung

Der untersuchte Datensatz basiert auf einer umfangreichen Befragung mit über 60 Fragen. Die praktische Nutzung dieser Daten ist jedoch mit mehreren Herausforderungen verbunden. Zum einen führt die Vielzahl erhobener Variablen zu hoher Dimensionalität, wodurch die Identifikation relevanter Einflussfaktoren erschwert wird. Zum anderen sind im Datensatz fehlende Werte enthalten, die nicht ohne Weiteres ignoriert werden können. Zusätzlich liegen qualitative Rückmeldungen der Mitarbeiter in Form unstrukturierter Textdaten vor, die sich mit klassischen statistischen Verfahren nicht direkt skalieren und enkodieren lassen.

1.2. Zielsetzung

Ziel dieser Arbeit ist es, den Datensatz durch geeignete Preprocessing- und Analyseverfahren so aufzubereiten, dass er als valide Entscheidungsgrundlage für die HR-Abteilung dient. Im Mittelpunkt steht dabei die Transformation komplexer Rohdaten in interpretierbare und relevante Informationen.

1.3. Vorgehensweise

Die methodische Struktur der Arbeit folgt einem klassischen Data-Science-Workflow, der sich in sechs aufeinanderfolgende Phasen gliedert. Den Ausgangspunkt bildet die Explorative Datenanalyse, um ein Verständnis für Verteilungen und Korrelationen zu gewinnen. Darauf folgt die Datenbereinigung, insbesondere die Imputation fehlender Werte. Im Schritt des Feature Engineering werden relevante Merkmale ausgewählt und gegebenenfalls neue Features generiert. Der Kern der Analyse besteht aus der Dimensionsreduktion, um den Datenraum zu komprimieren, gefolgt vom Clustering, um Muster zu segmentieren. Abschließend erfolgt die Interpretation der Cluster und die Visualisierung der Ergebnisse, um konkrete Handlungsempfehlungen für das Gesundheitsmanagement zu formulieren.

2. Datenbeschreibung und Explorative Datenanalyse

Im Rahmen der explorativen Datenanalyse (EDA) wurde der Datensatz hinsichtlich seiner Struktur, Verteilungen, fehlenden Werte sowie möglicher Inkonsistenzen untersucht. Dabei wurden zentrale Merkmale betrachtet und erste Muster identifiziert, die auf relevante Einflussfaktoren psychischer Belastung hinweisen. Die Ergebnisse der EDA bilden die Grundlage für die anschließenden Schritte der Datenvorverarbeitung und des Feature Engineerings. Zur besseren Nachvollziehbarkeit wird der zugehörige Code in diesem Dokument über den folgenden Verweismechanismus referenziert: `[FILEINDEX_CODELINE]`. Ein Eintrag wie `[0_31]` verweist beispielsweise auf die Datei `0_explorative_analysis.ipynb` und die dortige Codezeile 31. Der Quellcode ist auf GitHub zu finden.

2.1. Herkunft und Struktur des Datensatzes

Der verwendete Datensatz stammt aus der folgenden Quelle: Kaggle. Er basiert auf der OSMI Mental Health in Tech Survey 2016, einer internationalen Befragung mit über 1.400 Teilnehmenden aus der IT- und Tech-Branche. Ziel der Umfrage ist es, Einstellungen gegenüber psychischer Gesundheit am Arbeitsplatz zu erfassen sowie die Prävalenz psychischer Erkrankungen unter Beschäftigten in technischen Berufen zu untersuchen.

Der Datensatz umfasst insgesamt 1433 Zeilen, die jeweils einen Teilnehmenden repräsentieren, sowie 63 Spalten, die unterschiedlichen Fragen entsprechen `[0_11]`. Die Datentypen setzen sich aus `int64`, `float64` und überwiegend `object`-Typen zusammen, wobei Letztere hauptsächlich Freitexteingaben enthalten `[0_9]`. Insgesamt weist der Datensatz einen vergleichsweise hohen Anteil an fehlenden Werten auf.

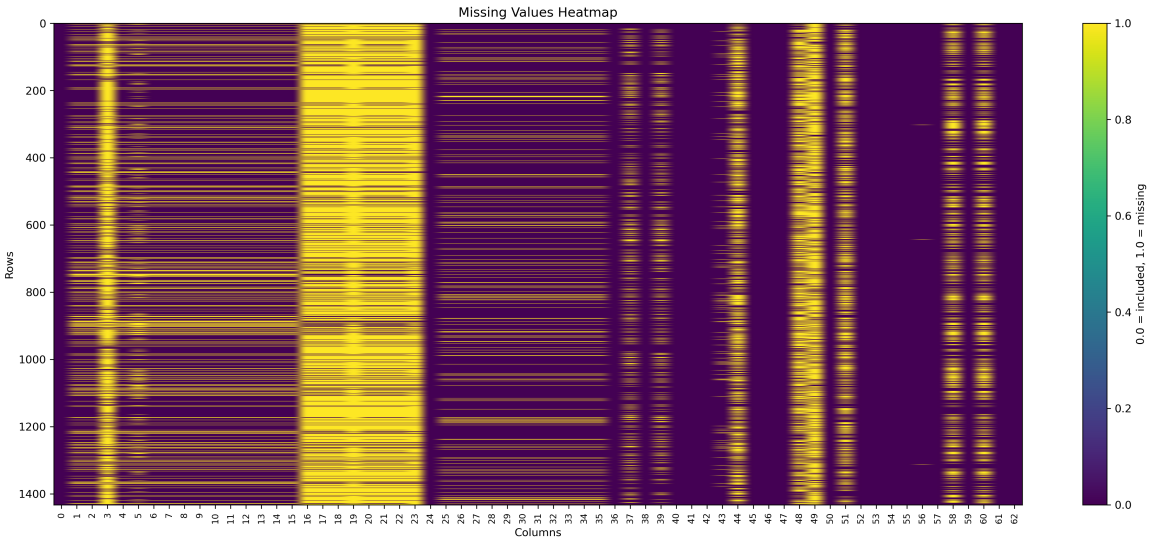

Abbildung 1.: Fehlende Werte des Datensatzes

Fehlende Einträge treten bereits in den Fragen 0 bis 36 regelmäßig auf, konzentrieren sich jedoch besonders stark im Bereich der Fragen 16 bis 24. Dabei handelt es sich überwiegend um sensible und persönliche Themen wie: “Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?” oder “If you have been diagnosed or treated for a mental health disorder, do you ever reveal this

to coworkers or employees?". Diese Fragen betreffen sehr private Aspekte der psychischen Gesundheit. Obwohl sie für die Fragestellung der Arbeit hohe Relevanz aufweisen, ist es nachvollziehbar, dass viele Teilnehmende sie unbeantwortet ließen.

2.2. Erste deskriptive Analysen

Im ersten Schritt wurden grundlegende demografische Variablen analysiert, darunter Geschlecht [0_14], Alter [0_15], Wohnland [0_40] und Arbeitsland [0_41]. Da die Variable Gender keine vordefinierten Antwortoptionen enthielt, konnten Teilnehmende beliebige Freitextangaben machen. Dies führte zu zahlreichen Varianten (z. B. Female, female, f) sowie zu unbrauchbaren Einträgen wie "Dudeöder "mail". Für eine konsistente Auswertung war daher eine nachträgliche Bereinigung und Zusammenführung der Kategorien erforderlich.

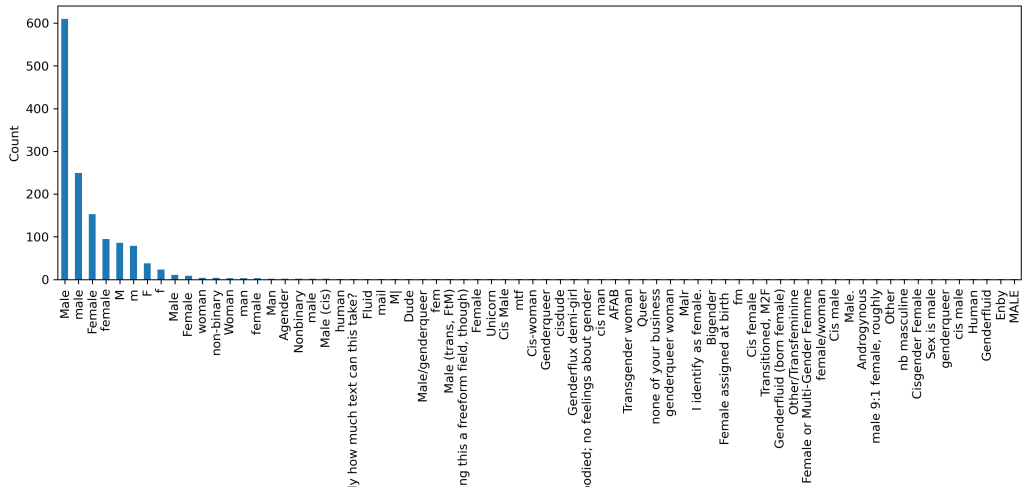

Abbildung 2.: Eingabe des Geschlechts

Bei der Variable Age traten deutliche Ausreißer auf, unter anderem Werte von über 100 bzw. sogar 300 Jahren [0_16]. Diese wurden im Rahmen der Datenbereinigung gesondert betrachtet.

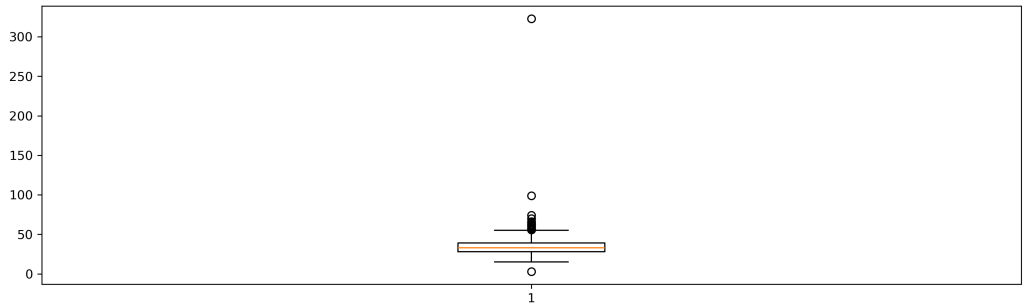

Abbildung 3.: Ausreißer beim Alter

Die Angaben zu Wohn- und Arbeitsland zeigen eine stabile und plausible Verteilung: Der Großteil der Befragten stammt aus den USA und dem Vereinigten Königreich. Zudem arbeiten die meisten Teilnehmenden in dem Land, in dem sie auch wohnen – 1407 Personen gaben an, im Wohnland zu arbeiten, während 26 Personen im Ausland tätig sind [0_23].

3. Datenvorverarbeitung

3.1. Umgang mit fehlenden Werten

Zu Beginn wurden alle offenen Freitextfragen entfernt, deren Inhalte für Clustering schwer interpretierbar sind (z. B. “Why or why not?”) sowie Fragen, die sich ausschließlich auf vorherige Antworten beziehen, etwa “What US state do you work in/live in?” [1_4]. Im Anschluss wurden fehlende Werte systematisch behandelt. Zunächst wurden alle Befragten gelöscht, die mehr als 40% der Fragen unbeantwortet ließen [1_5]. Anschließend wurden alle Fragen (Spalten) entfernt, deren Missing Ratio ebenfalls über 40% lag [1_7].

Für die verbleibenden Teilnehmenden mit weniger als 25% fehlenden Werten erfolgte eine Imputation [1_9]. Insgesamt enthielt der Datensatz 1014 Fragen, von denen jedoch nur diejenigen berücksichtigt wurden, die tatsächlich fehlende Werte aufwiesen. Für die Gruppe der Befragten mit einer Missing Ratio unter 0.25 wurden dabei lediglich drei Fragen identifiziert, die eine Imputation benötigten [1_12]:

- Frage 4: Antwortmöglichkeiten *no / not sure / yes*. Fehlende Werte wurden einheitlich mit *not sure* imputiert.
- Frage 32: Antwortmöglichkeiten *no / maybe-notsure / yes I observed / yes I experienced*. Fehlende Werte wurden auf *maybe-notsure* gesetzt.
- Frage 41: Ursprünglich drei Kategorien *male, female, others*. Da lediglich drei Werte fehlten, wurden diese der Kategorie *others* zugewiesen.

Die Vereinheitlichung der verschiedenen kategorischen Antworten für das Geschlecht wird im folgenden Kapitel detailliert beschrieben.

3.2. Ausreißer und Werte vereinheitlichen

Da viele Fragen als Freitexteingaben formuliert waren, traten zahlreiche uneinheitliche oder informelle Antworten auf, etwa “f”, “cis man” oder “none of your business”. Bevor fehlende Werte imputiert werden konnten, war daher eine umfassende Vereinheitlichung der Kategorien erforderlich. Dies erfolgte über ein regelbasiertes Mapping: Für die Geschlechtsangabe wurden beispielsweise Schlüsselwörter definiert (z. B. male, m, man). Enthielt eine Antwort eines dieser Keywords, wurde sie der Kategorie Male zugeordnet; entsprechend wurde mit Begriffen rund um Female verfahren. Alle übrigen Eingaben wurden der Kategorie Others zugewiesen. Die Umsetzung erfolgte mithilfe regulärer Ausdrücke (Regex) [1_10].

Für die Variable Age wurden offensichtliche Ausreißer ausgeschlossen. Teilnehmende unter 17 oder über 67 Jahren wurden als nicht plausibel eingestuft und entfernt. Anschließend wurden die validen Altersangaben in fünf Gruppen kategorisiert: *17–25*, *26–35*, *36–45*, *46–55* und *56–67* [1_14].

Die Variablen Wohnland und Arbeitsland wurden zu den zehn am häufigsten vorkommenden Ländern zusammengefasst. Alle weiteren Länder wurden in der Kategorie Others gebündelt [1_15].

Bei den Jobrollen zeigte sich eine besonders große Heterogenität, da viele Befragte mehrere Rollen gleichzeitig angaben und die Bezeichnungen stark variierten. Daher wurden alle Angaben in übergeordnete Hauptgruppen überführt: *Management/Lead*, *Developer*, *DevOps*, *Product Design*, *Data & Analytics*, *HR/Admin*, *Community*

und *Other*. Da einzelne Personen mehrere Rollen nannten, wurde zusätzlich ein Prioritätensystem eingeführt (z. B. Lead = 1, Developer = 2, . . . , Community = 7, Other = 99), um eine eindeutige Zuordnung zu gewährleisten [1_16].

3.3. Kodierung und Transformation der Merkmale

Nachdem die fehlenden Werte bereinigt und kategoriale Angaben vereinheitlicht wurden, konnte die eigentliche Merkmalskodierung durchgeführt werden. Je nach Skalenniveau der Variablen wurden unterschiedliche Verfahren angewendet

Binäre und nominale Merkmale wurden mittels One-Hot-Encoding transformiert. Vor der Kodierung wurden alle Fragen manuell identifiziert, deren Antwortmöglichkeiten ausschließlich binär oder nominal ausgeprägt sind. Nur diese Spalten wurden anschließend in den One-Hot-Encoder übergeben. Dies verhindert, dass ordinale oder numerische Variablen fälschlicherweise als nominal behandelt werden. Der Encoder wurde auf den relevanten Spalten gefittet und auf die Daten angewendet [1_18].

Alle verbleibenden kategorialen Merkmale wurden als ordinal betrachtet und mithilfe des *OrdinalEncoder* kodiert. Dabei wurde die Reihenfolge der Kategorien anhand ihrer tatsächlichen semantischen Struktur bzw. durch eine sinnvolle numerische Reihenfolge festgelegt. Antwortoptionen wie “I don’t know”, “Not applicable to me” oder “Unsure” wurden bewusst nicht entfernt, sondern stets als letzte Kategorie eingeordnet, sodass Modelle diese Werte eindeutig als separate Kategorie erkennen können. Diese Werte wurden nicht gelöscht, da sie von vielen Befragten gewählt wurden und potenziell wichtige Informationen enthalten [1_19].

Die Ergebnisse beider Kodierschritte wurden in einem neuen DataFrame data_encoded zusammengeführt, der alle transformierten Merkmale (One-Hot-Variablen und ordinal kodierte Spalten) umfasst. Der vollständig vorverarbeitete Datensatz wurde anschließend als data_preprocessed.csv gespeichert, um eine reproduzierbare Weiterverarbeitung in den nächsten Analyse- und Clustering-Schritten zu gewährleisten [1_20].

4. Feature Engineering

4.1. Feature Selection

Für Unsupervised Learning eignen sich insbesondere der Variance Threshold sowie die Korrelationsmatrix (scikit-learn developers, 2025), da beide Verfahren ohne Zielvariable funktionieren und helfen redundante oder nicht-informative Merkmale aus dem Datensatz zu entfernen. Der Variance Threshold identifiziert Merkmale ohne Varianz, während die Korrelationsmatrix hoch korrelierte Feature-Paare aufdeckt, deren Informationen sich stark überschneiden.

4.1.1. Variance Threshold

Zunächst wurde ein Variance Threshold von 0,01 angewendet. Dabei zeigte sich, dass bei zwei Merkmalen jeweils die ”NeinVariante (_0) bzw. ”JaVariante (_1) derselben Frage, alle Beobachtungen denselben Wert aufwiesen. Konkret betraf dies die Fragen ”Are you self-employed?” und ”Do you have previous employers?”, bei denen ausschließlich eine der beiden Kategorien vorkam. Da bei solchen Merkmalen die zweite Kategorie redundant ist, wurden diese Merkmale entfernt `[2_58]`.

4.1.2. Korrelationsmatrix

Im nächsten Schritt wurde eine Korrelationsmatrix berechnet, um stark korrelierte Merkmale zu erkennen `[2_59]`. Merkmale mit einer Korrelation von 1.00 wurden als redundant betrachtet. Dabei zeigte sich, dass bestimmte One-Hot-Encoder-Paare vollständig redundant sind: ”Is your employer primarily a tech company/organization?_0.0” und ”Is your employer primarily a tech company/organization?_1.0” hatten eine Korrelation von 1.00. Da es sich um eine ursprünglich binäre Variable handelt, reicht es aus, eine der beiden Spalten beizubehalten, ohne Informationsverlust. Das gleiche gilt für alle andere Merkmals-Paare, deren Korrelation 1.00 beträgt.

4.1.3. Wohnland vs. Arbeitsland

Die Analyse der Korrelationen zwischen Wohnland und Arbeitsland (beide ebenfalls One-Hot-enkodiert) zeigte für alle Paare extrem hohe Korrelationen, mit einem Minimum von 0.96 `[2_60]`. Da die überwiegende Mehrheit der Befragten in dem Land arbeitet, in dem sie auch wohnen, enthalten diese Merkmale praktisch die gleiche Information. Um Redundanz zu vermeiden, wurden daher alle Wohnland-Merkmale entfernt, während die Arbeitsland-Merkmale beibehalten wurden.

4.2. Methoden der Dimensionsreduktion

Dimensionsreduktion vereinfacht komplexe, hochdimensionale Datensätze, indem irrelevante oder redundante Merkmale entfernt werden. Das reduziert Rechenaufwand und erleichtert Analyse und Visualisierung, ohne die wichtigsten Informationen zu verlieren. Sie hilft außerdem, Overfitting und den „Curse of Dimensionality“ zu mindern, sodass Modelle robuster und generalisierungsfähiger werden (GmbH, 2025). Es wurden zwei

unterschiedliche Ansätze verwendet und verglichen, um denjenigen mit der besten Balance aus Interpretierbarkeit und Modellqualität zu identifizieren.

4.2.1. ANSATZ A: PCA

Zunächst wurde Principal Component Analysis verwendet. PCA eignet sich insbesondere für diesen Datensatz, da sie varianzbasierend arbeitet und alle Merkmale bereits enkodiert sowie standardisiert wurden (IBM, 2025). MDS (Multidimensional Scaling) wurde nicht verwendet, da es distanzbasiert ist und die Wahl eines geeigneten Distanzmaßes bei hochdimensionalen Fragebogendaten problematisch ist (Groenen und Borg, 2013). Auch LLE (Locally Linear Embedding) wurde nicht eingesetzt, da dieses Verfahren annimmt, dass die Daten auf einer nichtlinearen Mannigfaltigkeit liegen und viele Beobachtungen pro lokaler Nachbarschaft benötigen (Roweis und Saul, 2000). Diese Voraussetzungen sind bei typischen Umfragedaten meist nicht gegeben. Die Bestimmung der Anzahl der Hauptkomponenten erfolgte wie folgt: zunächst wurde die kumulative erklärte Varianz der PCA analysiert [2_65]. Der sogenannte Ellenbogenpunkt lag bei etwa 10 Komponenten, jedoch deckten diese lediglich 32% der Gesamtvarianz ab. Erst bei 64 Hauptkomponenten wurde ein akzeptabler Wert von 85% kumulierter Varianz erreicht. Aus diesem Grund wurde K = 64 als Anzahl der Hauptkomponenten gewählt.

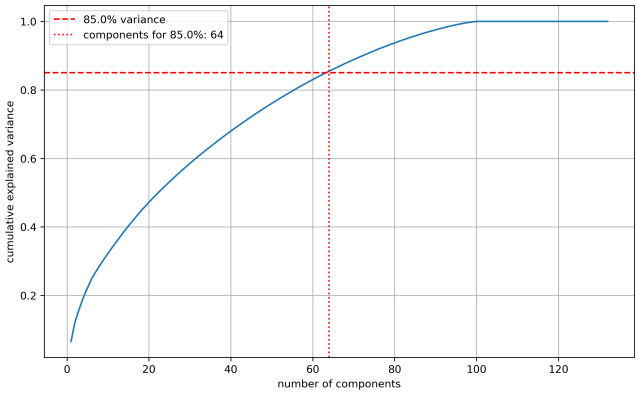

Abbildung 4.: Kumulative Erklärte Varianz für PCA

Die Wahl von 64 Hauptkomponenten führte allerdings zu zwei wesentlichen Problemen. Als Erstes erschwert eine so hohe Anzahl and Komponenten die inhaltliche Interpretation erheblich. Und als Zweites führte diese Methode zu schwachen Clustermetriken. Bei anschließenden Clustering-Versuchen zeigten sich ein sehr hoher BIC und extrem niedrieger Silhouettenwert (höchster Wert: 0.044 bei K = 2 [3_65])

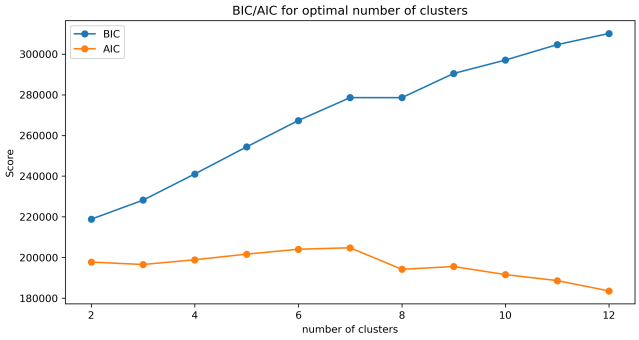

Abbildung 5.: BIC und AIC für PCA

Aufgrund der genannten Probleme wurde entschieden, nicht mit dem PCA-basierten Ansatz fortzufahren. Stattdessen wurde Ansatz B gewählt.

4.2.2. ANSATZ B: manuelle Feature Transformation

Beim zweiten Ansatz wurde der ursprüngliche, hochdimensionale Datensatz gezielt inhaltlich reduziert, indem thematisch zusammenhängende Fragen zu aussagekräftigen Merkmalen zusammengeführt wurden `[2_70]`. Dazu wurden Antworten inhaltlich gewichtet, zu neuen Feature-Scores aggregiert und anschließend die ursprünglichen Einzelmerkmale entfernt. Insgesamt wurden fünf neue, gut interpretierbare Scores gebildet:
employer_support_score misst, wie stark der derzeitige Arbeitgeber mentale Gesundheit unterstützt (höherer Wert = stärkerer Support), **prev_employer_support_score** ist analog zum obigen Score, jedoch für den vorherigen Arbeitgeber, **openness_score** erfasst, wie offen der Befragte gegenüber einem neuen Arbeitgeber in Bezug auf mentale Gesundheit wäre, **perceived_stigma_score** misst, wie stark der Befragte die Meinung vertritt, dass Offenheit über mentale Gesundheit der Karriere oder dem Team schaden könnte (höherer Wert = stärker wahrgenommenes Stigma) und **mh_status_score** bewertet den subjektiven mentalen Gesundheitsstatus der Person. Für jeden Score wurde eine Bewertungslogik verwendet. Antworten, die gegen den Score sprechen bekamen negative Werte, Antworten, die für den Score sprechen bekamen positive Werte. Pro Score wurden alle zugehörigen Fragen aufsummiert und anschließend der Durchschnitt berechnet `[2_71]`. Die fünf neuen Features wurden anschließend zum Datensatz hinzugefügt und mit `StandardScaler` standardisiert `[2_73]`. Durch diese manuelle Merkmalskonstruktion ergaben sich mehrere Vorteile: Reduktion der Dimensionalität auf 35 Merkmale, deutlich bessere Interpretierbarkeit und verbesserte Clustering-Metriken (BIC 35000 für K = 2 und Silhouettenwert 0.29 für K = 2)

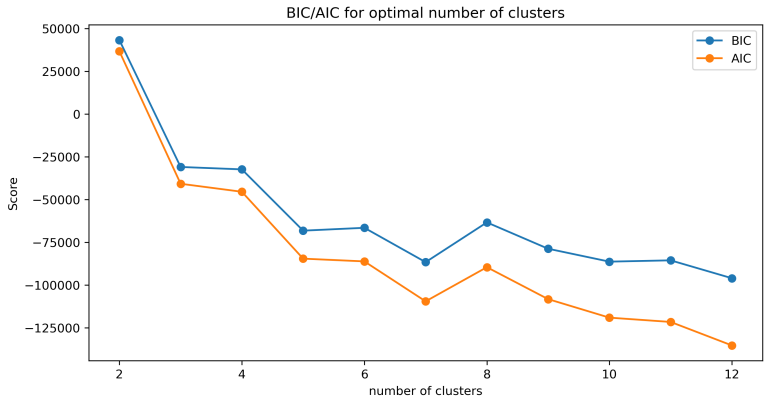

Abbildung 6.: BIC und AIC für manuell generierte Features

Der BIC sinkt mit steigender Clusteranzahl, also bevorzugt es höhere k-Werte. Der Silhouetten-Score zeigt dagegen eine klare Verschlechterung ab K > 4.

```
k = 2, Silhouette Score = 0.2951
k = 3, Silhouette Score = 0.2784
k = 4, Silhouette Score = 0.1305
k = 5, Silhouette Score = 0.1422
k = 6, Silhouette Score = 0.0936
k = 7, Silhouette Score = 0.0919
k = 8, Silhouette Score = 0.0878
k = 9, Silhouette Score = 0.0880
k = 10, Silhouette Score = 0.1081
```

Abbildung 7.: Silhouetten-Score für manuell generierte Features

Der BIC misst, wie gut ein Modell zu den Daten passt und bestraft unnötig viele Cluster. Niedrigere Werte sind besser, aber BIC neigt dazu, höhere Clusterzahlen zu bevorzugen. Der Silhouetten-Score misst, wie gut die Cluster voneinander getrennt sind. Für diese Aufgabe ist er wichtiger, da das Ziel klar interpretierbare Cluster sind. K = 2 hätte den besten Silhouetten-Score, aber der Unterschied zu K = 3 ist minimal. Da der BIC auch berücksichtigt wurde und K = 3 eine etwas sinnvollere Struktur bietet, wurde K = 3 gewählt.

5. Clustering

5.1. Auswahl geeigneter Methoden

Im Kontext der hochdimensionalen und heterogenen Umfragedaten erweist sich das Gaussian Mixture Model als am besten geeignetes Clustering-Verfahren. GMM erlaubt die Modellierung elliptischer und unterschiedlich großer Cluster und überwindet damit die Einschränkung von k-Means, das nur sphärische und gleich große Cluster zuverlässig erkennen kann. Gleichzeitig arbeitet GMM probabilistisch, was bei psychologischen Daten mit fließenden Übergängen zwischen Gruppen eine angemessene weiche Zuordnung ermöglicht (K, 2025). DBSCAN hingegen ist für hochdimensionale Daten problematisch, da seine Dichteparameter in solchen Räumen nicht mehr zuverlässig trennscharf wirken (GeeksforGeeks, 2025). Auch hierarchisch-agglomeratives Clustering ist aufgrund seiner hohen Rechenkomplexität und geringen Flexibilität bei komplexen Datensätzen weniger geeignet (Pullak, 2025). Im Vergleich dazu bietet GMM eine höhere Modellflexibilität, bessere Anpassungsfähigkeit an die tatsächliche Datenstruktur und eine verständliche Clusterzuordnung.

5.2. Bestimmung der Clusteranzahl

Wie zuvor beschrieben, wurden die geeigneten Clusterzahlen sowohl mittels PCA-basierter Analyse als auch durch manuell generierte Merkmalsselektionen überprüft. Beide Ansätze führten konsistent zu derselben Empfehlung. Auf dieser Grundlage wird die Clusteranzahl K = 3 gewählt.

5.3. Ergebnisse

Die Analyse identifiziert insgesamt drei Cluster, bestehend aus Cluster 0 (n = 102), Cluster 1 (n = 898) und Cluster 2 (n = 9). Im Folgenden werden die zentralen Merkmale der Cluster basierend auf den fünf generierten Kernvariablen dargestellt.

Tabelle 1.: Clustering von employer_support_score

Cluster 0	0.62	überdurchschnittlich
Cluster 1	0.54	überdurchschnittlich
Cluster 2	-1.2	stark unterdurchschnittlich

Cluster 2 weist eine deutlich unterdurchschnittliche wahrgenommene Arbeitgeberunterstützung auf, während Cluster 0 und 1 leicht überdurchschnittliche Werte zeigen.

Tabelle 2.: Clustering von prev_employer_support_score

Cluster 0	0.19	durchschnittlich
Cluster 1	0.89	überdurchschnittlich
Cluster 2	-1.1	stark unterdurchschnittlich

Cluster 2 berichtet sowohl früher als auch heute von klar negativen Erfahrungen, während Cluster 0 durchschnittliche und Cluster 1 überdurchschnittlich positive Erfahrungen aufweist.

Tabelle 3.: Clustering von openness_score

Cluster 0	-0.69	unterdurchschnittlich
Cluster 1	-0.46	unterdurchschnittlich
Cluster 2	1.1	stark überdurchschnittlich

Cluster 2 zeichnet sich durch eine stark überdurchschnittliche Offenheit aus, während die Cluster 0 und 1 jeweils unterdurchschnittliche Werte zeigen.

Tabelle 4.: Clustering von perceived_stigma_score

Cluster 0	-0.83	stark unterdurchschnittlich
Cluster 1	-0.29	unterdurchschnittlich
Cluster 2	1.1	stark überdurchschnittlich

Cluster 2 weist eine deutlich überdurchschnittliche Stigmatisierung auf. Cluster 0 und 1 hingegen empfinden ein geringes Stigma im beruflichen Kontext.

Tabelle 5.: Clustering von mh_status_score

Cluster 0	-1.1	stark unterdurchschnittlich
Cluster 1	0.36	leicht überdurchschnittlich
Cluster 2	0.77	überdurchschnittlich

Cluster 0 zeigt einen stark eingeschränkten mentalen Gesundheitszustand, während Cluster 2 überdurchschnittlich und Cluster 1 leicht überdurchschnittlich abschneidet. Es ergibt sich folgendes Clusterprofil:
Cluster 2: Offene, mental stabile Personen mit gleichzeitig geringer wahrgenommener Arbeitgeberunterstützung und hohem Stigmaerleben.
Cluster 0: Weniger offene Personen mit schwachem mentalem Gesundheitszustand, trotz moderater Arbeitgeberunterstützung.
Cluster 1: Vergleichsgruppe mit überwiegend durchschnittlichen bis leicht positiven Ausprägungen.

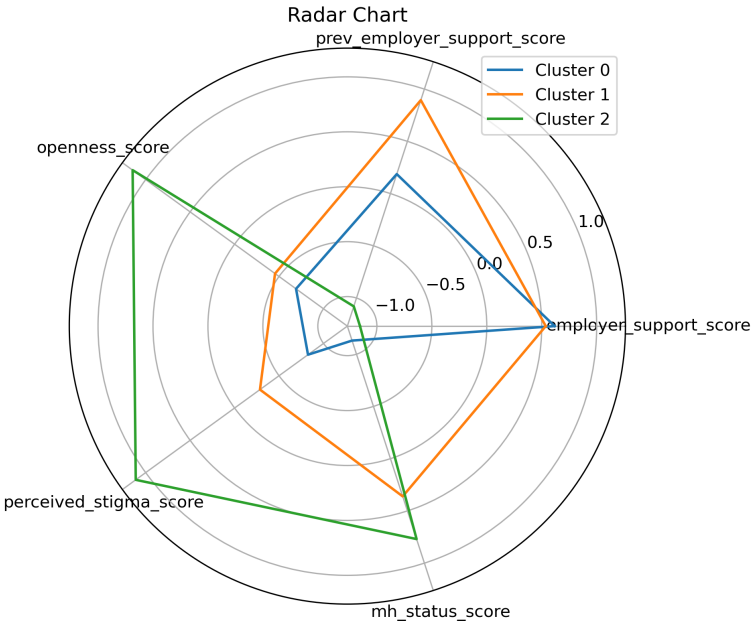

Abbildung 8.: Fünf selbst-generierte Features pro Cluster

Das Radar-Diagramm zeigt die drei Cluster im direkten Vergleich und macht deren Profilunterschiede über die fünf ausgewählten Features gut sichtbar. Es verdeutlicht damit die strukturellen Unterschiede zwischen den Clustern auf einen Blick.

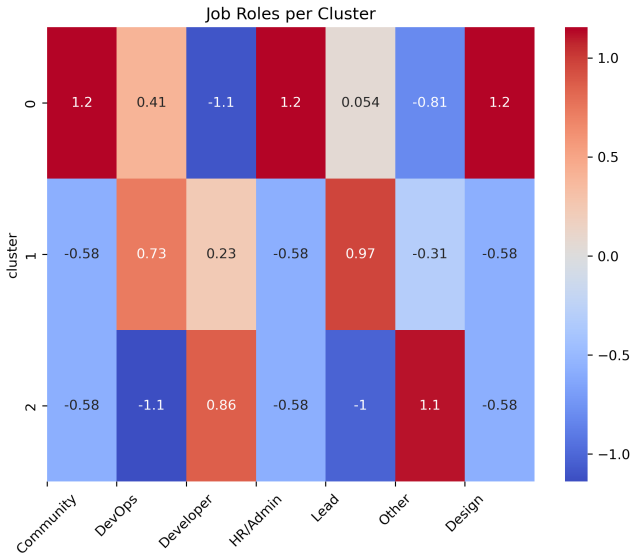

Abbildung 9.: Arbeitsstellen pro Cluster

Cluster 0 enthält vorwiegend Rollen aus Community/Advocacy, HR/Administration und Design, ergänzt durch einzelne DevOps-Positionen. Cluster 1 besteht hauptsächlich aus DevOps, einigen Entwicklern und Führungskräften. Cluster 2 umfasst überwiegend Developer sowie nicht näher spezifizierte Rollen.

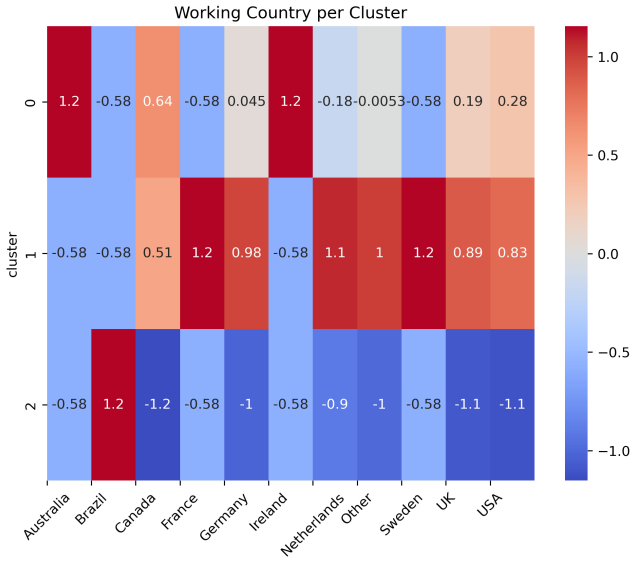

Abbildung 10.: Arbeitsländer pro Cluster

Cluster 0 ist überwiegend in Australien und Irland vertreten, gefolgt von Kanada, dem Vereinigten Königreich und den USA. Cluster 1 setzt sich hauptsächlich aus Personen aus Frankreich, Deutschland, den Niederlanden, Schweden, dem Vereinigten Königreich und den USA zusammen. Cluster 2 besteht nahezu ausschließlich aus Personen aus Brasilien.

Tabelle 6.: Clustering von Tech Company

Cluster 0	0.91	stark überdurchschnittlich
Cluster 1	0.17	durchschnittlich
Cluster 2	-1.07	stark unterdurchschnittlich

Cluster 0 arbeitet überwiegend in Tech-Unternehmen, während Cluster 2 klar unterdurchschnittlich im Tech-Sektor beschäftigt ist. Cluster 1 liegt hier im Durchschnittsbereich.

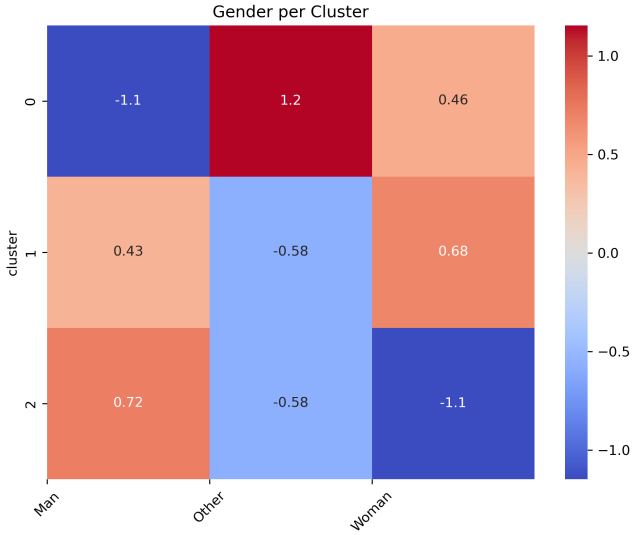

Abbildung 11.: Geschlecht pro Cluster

Cluster 0 weist eine hohe Konzentration an Frauen und nicht-binären Personen auf. Cluster 2 besteht überwiegend aus Männern, während Cluster 1 eine nahezu ausgeglichene Geschlechterverteilung aufweist.

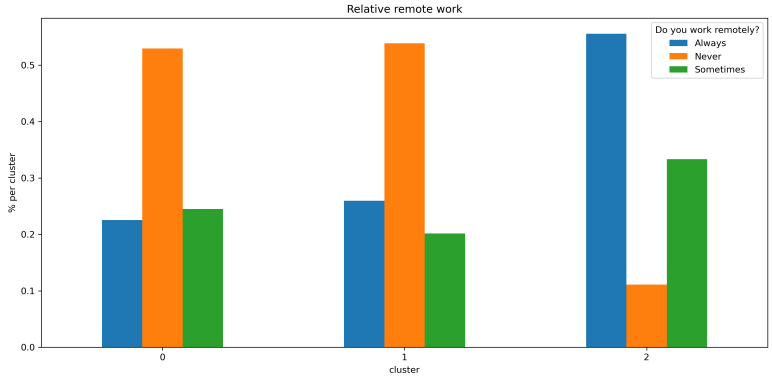

Abbildung 12.: Remote Work pro Cluster (relativ)

Cluster 0 und 1 berichten überwiegend, nie remote zu arbeiten, während Cluster 2 deutlich häufiger im Home-Office arbeitet.

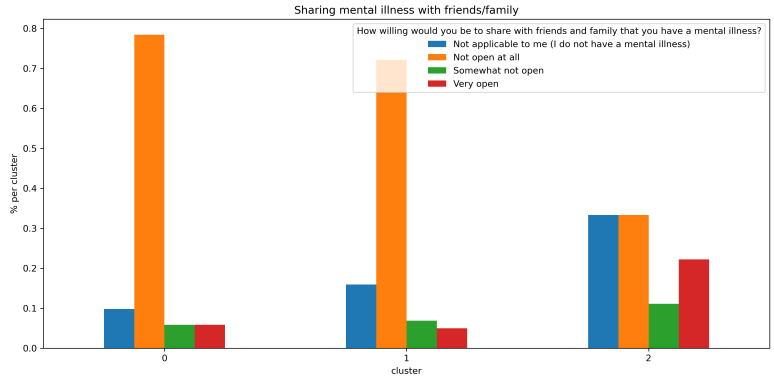

Abbildung 13.: Über MH-Problemen mit Familie teilen pro Cluster (relativ)

Cluster 0 und 1 würden potenzielle mentale Probleme selten mit Familie oder Freunden teilen, während Cluster 2 ein ausgeglicheneres Kommunikationsverhalten zeigt.

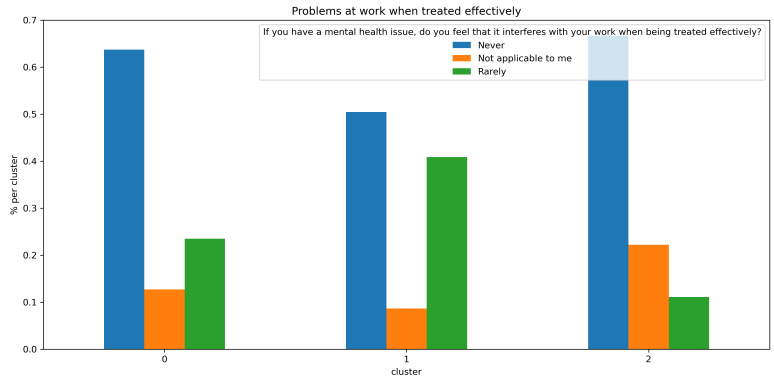

Abbildung 14.: Schwierigkeiten in der Arbeit bei guter Behandlung

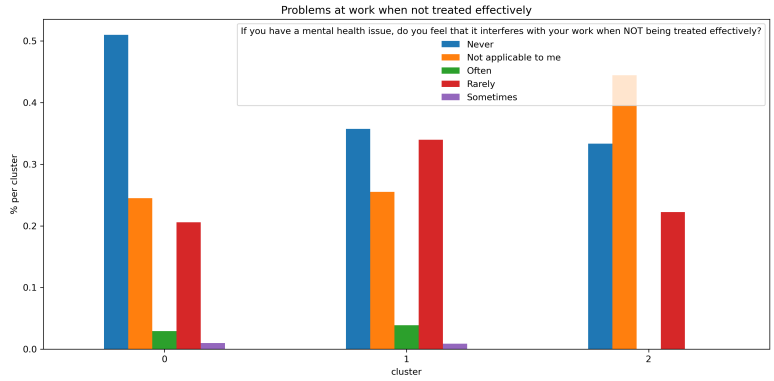

Abbildung 15.: Schwierigkeiten in der Arbeit bei schlechter Behandlung

Die Ergebnisse zeigen, dass eine ineffektive Behandlung mentaler Probleme am Arbeitsplatz deutlich häufiger zu arbeitsbezogenen Schwierigkeiten führt als eine effektive Unterstützung.

5.4. Übertragung auf HR-Kontext

Die Ergebnisse der Clusteranalyse zeigen drei klar voneinander abgegrenzte Gruppen von Mitarbeitenden, die sich sowohl in ihren Belastungsfaktoren als auch in ihren Ressourcen deutlich unterscheiden. Cluster 0 umfasst

Personen, deren Belastungen vor allem in individuellen Faktoren verankert sind. Sie weisen einen gering ausgeprägten mentalen Gesundheitsstatus, eine niedrige Offenheit im Umgang mit psychischen Belastungen sowie ein reduziertes soziales Unterstützungsverhalten auf. Charakteristisch sind zudem Tätigkeiten im HR-, Design- oder administrativen Bereich sowie eine überwiegend stationäre Arbeitsweise. Diese Ausprägungen deuten auf einen Bedarf an gezielten Maßnahmen zur individuellen Unterstützung hin, insbesondere durch intensivere Mental-Health-Angebote, eine Stärkung der Führungskompetenzen zur Erkennenung früherer Warnsignale, einen erleichterten Zugang zu Unterstützungsprogrammen sowie eine Erhöhung flexibler Arbeitsoptionen.

Cluster 1 bildet eine stabile Vergleichsgruppe mit leicht überdurchschnittlichen Werten in den zentralen Variablen und einer ausgewogenen demografischen Verteilung. Die hier vertretenen Rollen stammen überwiegend aus DevOps und Engineering. Für diese Gruppe steht weniger der Bedarf nach neuen Interventionen im Vordergrund, sondern vielmehr die Stabilisierung der bestehenden Strukturen sowie ein kontinuierliches Monitoring möglicher Belastungsentwicklungen. Darüber hinaus bietet dieses Cluster Potenzial für den Transfer erfolgreicher Praktiken auf die beiden anderen Gruppen.

Cluster 2 weist demgegenüber eine hohe Offenheit und psychische Stabilität auf. Die Mitglieder dieser Gruppe erleben jedoch gleichzeitig ein ausgeprägtes berufliches Stigma und eine vergleichsweise geringe Arbeitgeberunterstützung. Diese Kombination ist bemerkenswert und weist auf systemische Defizite in Unternehmenskultur, Kommunikation oder vorhandenen Unterstützungsstrukturen hin. Die ausgeprägte psychische Stabilität könnte zudem mit der länderspezifischen Zusammensetzung dieses Clusters zusammenhängen, da der nahezu vollständige Fokus auf Brasilien darauf hindeutet, dass kulturell verankerte Verhalten und Denkweisen eine Rolle spielen könnten. Trotz dieser individuellen Ressourcen fehlen verlässliche organisationale Rahmenbedingungen, die einen offenen Umgang mit mentalen Themen ermöglichen. Folglich liegt der Schwerpunkt des Bedarfs auf dem Ausbau wirksamer Mental-Health-Policies sowie der Etablierung klarer Prozesse für Gespräche über psychische Gesundheit.

Die Gegenüberstellung der drei Cluster verdeutlicht zudem, dass das Vorhandensein unternehmerischer Unterstützungsprogramme allein keine Garantie für deren Wirksamkeit darstellt. In Cluster 0 bestehen Unterstützungsangebote, dennoch weisen die Mitarbeitenden dort die schwächste mentale Gesundheit und die geringste Offenheit auf. Cluster 2 hingegen verfügt über kaum institutionelle Unterstützung, zeigt jedoch eine vergleichsweise stabile psychische Lage. Diese Konstellation legt nahe, dass Programme nur dann Wirkung entfalten können, wenn sie in eine passende Kultur eingebettet sind und tatsächlich zugänglich und bedarfsgerecht genutzt werden.

6. Schluss

Das gewählte methodische Vorgehen hat sich insgesamt als positiv erwiesen. Besonders gut funktionierten die Datenaufbereitung sowie das Clustering. Die klare Strukturierung entlang des klassischen Data-Science-Workflows ermöglichte eine nachvollziehbare Analyse und eine solide Grundlage für die spätere Interpretation. Auch die Integration neuer Merkmale erwies sich als hilfreich, um ein grobes Gesamtbild zu erhalten. Herausfordernder war hingegen der Umgang mit der hohen Dimensionalität des Datensatzes und den zahlreichen fehlenden Werten. Trotz sorgfältiger Imputationsmethoden bleibt die Gefahr bestehen, dass bestimmte Muster dadurch abgeschwächt oder verstärkt wurden. Alternative Ansätze wie robustere Feature-Selection-Methoden, nichtlineare Dimensionsreduktion oder andere Clusterverfahren hätten zusätzliche Perspektiven eröffnet und möglicherweise feinere Strukturen sichtbar gemacht.

Die Analyse unterliegt mehreren Einschränkungen. Erstens ist die Qualität der Umfragedaten heterogen. Subjektive Selbsteinschätzungen sind anfällig für Verzerrungen, und der Anteil fehlender Werte erschwert eine zuverlässige Interpretation. Zweitens ist die Generalisierbarkeit der Ergebnisse begrenzt, da die Stichprobe in einigen Clustern stark von spezifischen Regionen oder Berufsgruppen geprägt ist. Drittens wurden einige potenziell relevante Faktoren nicht berücksichtigt, darunter organisational-historische Entwicklungen oder externe Belastungsfaktoren außerhalb des Arbeitsplatzes. Diese könnten zukünftige Analysen weiter differenzieren.

Die Clusteranalyse identifizierte drei klar voneinander abgegrenzte Mitarbeitendengruppen. Cluster 0 weist ein personenzentriertes Belastungsprofil mit niedriger psychischer Gesundheit, geringer Offenheit und eingeschränkter sozialer Unterstützung auf. Cluster 1 bildet eine stabile Vergleichsgruppe mit leicht positiven Ausprägungen in nahezu allen Kernvariablen sowie einer ausgewogenen demografischen Struktur. Cluster 2 zeigt ein organisationszentriertes Belastungsmuster, geprägt durch hohe Offenheit und psychische Stabilität, jedoch gleichzeitig sehr schwache Arbeitgeberunterstützung und stark ausgeprägtes berufliches Stigma. Auffällig ist, dass diese psychische Stabilität möglicherweise mit der Länderzusammensetzung zusammenhängt, da dieses Cluster nahezu ausschließlich aus Personen aus Brasilien besteht. Besonders problematisch ist die Feststellung, dass vorhandene Unterstützungsprogramme nicht automatisch wirksam sind. Cluster 0 erhält zwar prinzipiell betriebliche Angebote, befindet sich jedoch in der schlechtesten mentalen Lage. Cluster 2 hingegen verfügt kaum über Unterstützung, zeigt aber dennoch hohe psychische Stabilität.

Für zukünftige Arbeiten bietet sich die Integration weiterer Datenquellen wie Leistungsmetriken oder Teamstrukturen an, um ein übersichtlicheres Bild der Belastungslagen zu erhalten. Ebenso wäre ein kontinuierliches Monitoring sinnvoll, um die Wirksamkeit spezifischer Interventionen zu prüfen und Veränderungen im Zeitverlauf sichtbar zu machen. Darüber hinaus eröffnet der Einsatz fortgeschrittener ML-Modelle hohes Potenzial für ein präventiv ausgerichtetes Gesundheitsmanagement.

Literaturverzeichnis

GeeksforGeeks. (2025). Choosing the right clustering algorithm for your dataset [Zugriff am 08.12.2025]. https://www.geeksforgeeks.org/data-science/choosing-the-right-clustering-algorithm-for-your-dataset/

GmbH, E. (2025). Dimensionsreduktion: Wie erleichtert sie große Datenmengen? [Zugriff am 05.12.2025]. https://evoluce.de/dimensionsreduktion/

Groenen, P. J., & Borg, I. (2013). *The Past, Present, and Future of Multidimensional Scaling* (Econometric Institute Report Nr. EI2013-07) (s. 3). Econometric Institute, Erasmus University Rotterdam. https://repub.eur.nl/pub/39177/EI2013-07.pdf

IBM. (2025). Was ist die Principal Component Analysis (PCA)? [Zugriff am 08.12.2025]. https://www.ibm.com/de-de/think/topics/principal-component-analysis

K, A. (2025). Clustering Algorithms in Machine Learning: Types, Comparison & Accuracy [Zugriff am 08.12.2025]. https://learninglabb.com/clustering-algorithms-in-machine-learning/

Pullak, K. (2025). Guide to Clustering Algorithms: Strengths, Weaknesses and Evaluation [Zugriff am 08.12.2025]. https://krishnapullak.medium.com/guide-to-clustering-algorithms-strengths-weaknesses-and-evaluation-5285a75ea902

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

scikit-learn developers. (2025). Feature selection — scikit-learn documentation [Zugriff am 03.12.2025]. https://scikit-learn.org/stable/modules/feature_selection.html