Projekt: Maschinelles Lernen - Unsupervised Learning und Feature Engineering DLBDSMLUSL01_D

Fallstudie

Studiengang: Angewandte Künstliche Intelligenz

Sven Behrens

Matrikelnummer: 42303511

Prof. Dr. Christian Müller-Kett

10. Januar 2026

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ADHD Attention Deficit Hyperactivity Disorder

AFAB Assigned Female At Birth

GMM Gaussian Mixture Models

LLE Locally Linear Embedding

MDS Multidimensional Scaling

OCD Obsessive-Compulsive Disorder

OSMI Open Sourcing Mental Illness

PCA Principal Component Analysis

PTSD Post-traumatic Stress Disorder

t-SNE t-Distributed Stochastic Neighbor Embedding

UMAP Uniform Manifold Approximation and Projection

1 Einleitung

Immer mehr Menschen erkranken an psychischen Erkrankungen (World Health Organization, 2025). Auch in technologieorientierten Berufen rücken psychische Belastungen am Arbeitsplatz verstärkt in den Fokus von Unternehmen und Forschung. Die systematische Analyse von Umfragedaten zur psychischen Gesundheit stellt dabei eine zentrale Herausforderung dar, deren Bewältigung maßgeblich zur Entwicklung zielgerichteter Präventionsprogramme und zur Verbesserung der Arbeitsbedingungen beitragen kann. Vor diesem Hintergrund wurde im Rahmen des Moduls „Projekt: Maschinelles Lernen – Unsupervised Learning und Feature Engineeringän der IU Internationalen Hochschule eine umfassende Clusteranalyse von Umfragedaten zur psychischen Gesundheit in der Technologiebranche durchgeführt.

Das primäre Projektziel bestand in der Kategorisierung von Umfrageteilnehmenden anhand ihrer Antworten zu psychischen Belastungen, Arbeitgeberunterstützung und Stigmatisierungserfahrungen mittels unüberwachter Lernverfahren. Die zentrale Forschungsfrage konzentrierte sich darauf, wie durch den Einsatz verschiedener Dimensionsreduktions- und Clustering-Methoden aussagekräftige Teilnehmergruppen identifiziert werden können, die als Grundlage für gezielte Interventionsmaßnahmen der Personalabteilung dienen. Besondere Aufmerksamkeit galt dabei der Interpretierbarkeit der Ergebnisse sowie der Reduktion der hohen Dimensionalität des Datensatzes bei gleichzeitiger Beibehaltung der wesentlichen Informationsstruktur.

Die Datenbasis bildete der OSMI Mental Health in Tech Survey 2016 (Open Sourcing Mental Illness, 2016), der auf Kaggle frei verfügbar ist und Antworten von 1.433 Beschäftigten aus technologieorientierten Unternehmen umfasst. Der Datensatz enthält 63 Fragen zu Themen wie diagnostizierte psychische Erkrankungen, Einstellungen gegenüber psychischer Gesundheit am Arbeitsplatz, wahrgenommene Arbeitgeberunterstützung, Stigmatisierungserfahrungen sowie demografische Merkmale. Die Herausforderungen bei der Arbeit mit diesen Daten lagen insbesondere in der hohen Anzahl fehlender Werte, nicht standardisierten Texteingaben z.B bei Geschlechtsangaben, sowie der Notwendigkeit einer geeigneten Kodierung kategorischer Variablen für maschinelle Lernalgorithmen.

Die methodische Vorgehensweise gliederte sich in mehrere aufeinander aufbauende Phasen. Nach einer initialen explorativen Datenanalyse erfolgte eine umfassende Datenvorverarbeitung, die Bereinigung, Normalisierung, Imputation fehlender Werte sowie Feature Engineering umfasste. Anschließend wurden verschiedene Dimensionsreduktionsmethoden, darunter Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Multidimensional Scaling (MDS) und Locally Linear Embedding (LLE), systematisch evaluiert und verglichen. Für die Clusteranalyse kamen K-Means, Gaussian Mixture Models (GMM) sowie hierarchisches Clustering zum Einsatz, wobei die optimale Clusteranzahl durch verschiedene Evaluationsmetriken bestimmt wurde.

Der gewählte Ansatz zeichnet sich durch seine iterative Vorgehensweise aus, bei der in drei aufeinander

aufbauenden Durchläufen die Datenvorverarbeitung und Modellierung schrittweise verfeinert wurden. Durch die Kombination verschiedener Methoden und die sorgfältige Interpretation der Ergebnisse wurde ein Analysesystem geschaffen, das nicht nur statistisch fundierte Cluster identifiziert, sondern auch praxisrelevante Erkenntnisse für die Gestaltung von Präventionsprogrammen liefert.

Die vorliegende Fallstudie dokumentiert systematisch den gesamten Analyseprozess. Nach dieser Einleitung folgt die detaillierte Beschreibung der Datenbasis sowie der durchgeführten Vorverarbeitungsschritte. Anschließend werden die angewandten Dimensionsreduktions- und Clustering-Methoden erläutert und deren Ergebnisse ausführlich dargestellt. Abschließend werden die gewonnenen Erkenntnisse in einem Fazit zusammengefasst und kritisch reflektiert. Da die Analyse sensibler Gesundheitsdaten besondere Sorgfalt erfordert, werden zudem ethische und gesellschaftliche Aspekte wie Datenschutz, Stigmatisierungsrisiken und der verantwortungsvolle Umgang mit den Ergebnissen eingehend diskutiert.

2 Hauptteil

2.1 Projektumgebung

Zu Beginn des Projekts wurde ein GitHub-Repository[1] angelegt, um eine nachvollziehbare Versionsverwaltung zu gewährleisten und bei Bedarf auf frühere Entwicklungsstände zurückgreifen zu können. Anschließend wurde eine grundlegende Verzeichnisstruktur erstellt, die separate Ordner für Notebooks, Daten, Visualisierungen und den Bericht umfasst. Für die Entwicklung wurde mit venv eine virtuelle Python-Umgebung eingerichtet, in der die benötigten Bibliotheken wie NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn und UMAP-learn installiert wurden. Als Entwicklungsumgebung dienten Jupyter Notebooks. Abschließend wurde der OSMI Mental Health in Tech Survey 2016 Datensatz von Kaggle heruntergeladen und im Datenverzeichnis abgelegt.

2.2 Iterativer Analyseansatz

Die Analyse wurde in einem iterativen Prozess durchgeführt, der drei aufeinander aufbauende Durchläufe umfasste. Dieser Ansatz ermöglichte es, aus den Erkenntnissen jedes Durchlaufs systematisch zu lernen und die Methodik kontinuierlich zu verbessern. Jeder Durchlauf folgte dabei einem einheitlichen Ablauf bestehend aus Datenvorverarbeitung, Dimensionsreduktion, Clustering und anschließender Interpretation der Ergebnisse.

[1] https://github.com/svenb23/ML-UL-FE

2.3 Erster Durchlauf

2.3.1 Datenexploration und Vorverarbeitung

Der Datensatz umfasste 1.433 Teilnehmende und 63 Features, die verschiedene Aspekte der psychischen Gesundheit am Arbeitsplatz abdeckten. Die Features ließen sich thematisch in demografische Merkmale, Angaben zum Arbeitsumfeld, persönliche psychische Gesundheitssituation, Unterstützung durch aktuelle und frühere Arbeitgeber sowie Einstellungen und Stigmatisierungserfahrungen einteilen. Eine initiale Prüfung auf offensichtlich irrelevante Spalten ergab keine Kandidaten und wurde aufgrund mangelnden domänenspezifischen Wissens nicht weiter verfolgt.

Die Analyse der Datentypen zeigte eine deutliche Dominanz kategorischer Variablen: 56 der 63 Features lagen als Objekttyp vor, während lediglich vier ganzzahlige und drei Gleitkomma-Spalten existierten. Eine differenziertere Betrachtung der Datentypen ergab die in Table 1 dargestellte Verteilung.

Bei der Ausreißeranalyse wurde die Altersspalte untersucht, da diese als einzige numerische Variable anfällig für Fehleingaben war. Dabei wurden sechs auffällige Werte identifiziert: 3, 15, 17, 74, 99 und 323 Jahre. Die Werte 3, 99 und 323 wurden als offensichtliche Fehleingaben klassifiziert und aus dem Datensatz entfernt. Die Grenzfälle 15, 17 und 74 Jahre wurden beibehalten, da sie im Kontext der Technologiebranche als möglich eingestuft wurden.

Das Freitextfeld für Geschlechtsangaben führte zu 69 unterschiedlichen Einträgen. Diese reichten von Schreibvarianten wie „Male", „male", „Münd „Man"über Bezeichnungen wie „non-binaryünd „Agender"bis hin zu „Transgender womanöder „AFAB". Da eine differenzierte Betrachtung von biologischem und empfundenem Geschlecht über den Fokus dieser Analyse hinausgeht, wurden die Angaben pragmatisch auf Basis des bei Geburt zugewiesenen Geschlechts in drei Kategorien zusammengefasst: Male (1.055), Female (338) und Other (34).

Die Analyse fehlender Werte ergab, dass 44 der 63 Spalten Lücken aufwiesen (siehe Table 2). Der Anteil reichte von 0,2% bis zu 89,9% pro Spalte. Spalten mit mehr als 70% fehlenden Werten wurden entfernt, da eine sinnvolle Imputation bei diesem Anteil nicht mehr gewährleistet werden konnte. Ebenso wurden die US-Bundesstaaten-Spalten ausgeschlossen, da sie nur für amerikanische Teilnehmende relevant waren. Für Fragen zum aktuellen Arbeitgeber wurden fehlende Werte bei Selbstständigen mit „Not applicable"gefüllt, da diese Fragen für sie nicht zutrafen. Die verbleibenden fehlenden Werte in kategorischen Spalten wurden mit dem Modus imputiert.

Besondere Aufmerksamkeit erforderten Spalten mit Mehrfachnennungen, bei denen Teilnehmende mehrere Optionen auswählen konnten. Dies betraf insbesondere die diagnostizierten psychischen Erkrankungen sowie die Arbeitspositionen. Für diese Spalten wurden verschiedene Encoding-Strategien evaluiert: einfaches One-Hot-Encoding, Count-Encoding, Primärkategorie-Extraktion und gruppiertes One-Hot-

Encoding. Die Wahl fiel auf gruppiertes One-Hot-Encoding, das die Vorteile der Informationserhaltung mit einer Reduktion der Dimensionalität verbindet.

Die psychischen Erkrankungen wurden in vier semantische Gruppen zusammengefasst: angstbezogene Störungen (Anxiety, Obsessive-Compulsive Disorder (OCD), Post-traumatic Stress Disorder (PTSD), Stress), stimmungsbezogene Störungen (Depression, Bipolar), neurologische Entwicklungsstörungen (Attention Deficit Hyperactivity Disorder (ADHD)) sowie sonstige Erkrankungen (Substanzmissbrauch, Essstörungen, Persönlichkeits- und psychotische Störungen). Analog wurden die Arbeitspositionen in die Gruppen Developer, Operations, Leadership und Other unterteilt. Durch diese Gruppierung entstanden 12 neue binäre Spalten anstelle von geschätzten 44 Einzelkategorien.

Für ordinale Variablen mit natürlicher Rangordnung wurde eine numerische Kodierung gewählt, die die Reihenfolge erhält und sinnvolle Distanzberechnungen ermöglicht. Sechs Spalten wurden entsprechend transformiert: Unternehmensgröße (1–6, wobei 0 für Selbstständige), Schwierigkeit der Krankmeldung (1–5), Offenheit gegenüber Familie und Freunden (1–5), Remote-Arbeitsanteil (0–2) sowie die Beeinträchtigung der Arbeit durch psychische Probleme mit und ohne Behandlung (jeweils 0–3).

Die nominalen Variablen ohne natürliche Ordnung wurden mittels One-Hot-Encoding transformiert. Eine Ausnahme bildeten die Länderspalten mit 53 verschiedenen Ausprägungen: Aufgrund der starken Dominanz der USA (58,6% der Teilnehmenden) wurden diese binär kodiert (USA vs. Non-USA), um die Dimensionalität zu begrenzen.

Die beiden Freitextfelder hätten mittels Natural Language Processing weiterverarbeitet werden können, beispielsweise durch Sentimentanalyse (Behrens, 2025), Topic Modeling oder die Extraktion von Text-Embeddings. Da der Fokus dieser Analyse jedoch auf den strukturierten Daten lag und die Integration von Textfeatures die Komplexität erheblich erhöht hätte, wurden die Freitextfelder erst einmal aus dem Datensatz entfernt.

Nach Abschluss aller Vorverarbeitungsschritte wurde der transformierte Datensatz als CSV-Datei gespeichert, um ihn für die nachfolgenden Analyseschritte zur Verfügung zu stellen. Der finale Datensatz umfasste 1.427 Zeilen und 159 Spalten.

2.3.2 Dimensionsreduktion

Für die Dimensionsreduktion wurde der vorverarbeitete Datensatz geladen und zunächst mit dem StandardScaler normalisiert. Diese Standardisierung transformiert jede Variable auf einen Mittelwert von 0 und eine Standardabweichung von 1, was für distanzbasierte Methoden wie PCA und t-SNE essenziell ist, da ansonsten Variablen mit größeren Wertebereichen die Ergebnisse dominieren würden.

Die PCA wurde als erste Methode eingesetzt, da sie eine lineare, deterministische Transformation bietet,

die die Hauptvarianzrichtungen im Datensatz identifiziert. Ihre Vorteile liegen in der Interpretierbarkeit der Komponenten, der schnellen Berechnung und der Möglichkeit, die erklärte Varianz zu quantifizieren. Nachteilig ist, dass PCA nur lineare Zusammenhänge erfasst und bei komplexen, nicht-linearen Strukturen an ihre Grenzen stößt (IU Internationale Hochschule, 2025, S. 77–84). Die Analyse ergab, dass 49 Komponenten für 80% und 80 Komponenten für 95% der erklärten Varianz benötigt wurden. Ein Indikator für die hohe Komplexität des Datensatzes nach dem One-Hot-Encoding. Für die 2D-Visualisierung erklärten die ersten beiden Hauptkomponenten zusammen nur etwa 8% der Gesamtvarianz, was eine begrenzte Aussagekraft der Projektion impliziert. Dennoch ließen sich in der Projektion vier visuell unterscheidbare Cluster erkennen. Als Vorverarbeitungsschritt für rechenintensive Methoden wurde eine PCA mit 50 Komponenten durchgeführt, die etwa 80% der Varianz erhielt und gleichzeitig die Dimensionalität deutlich reduzierte.

MDS wurde als zweite Methode eingesetzt, um die paarweisen Distanzen zwischen den Datenpunkten in einer niedrigdimensionalen Darstellung zu erhalten. Der Vorteil von MDS liegt in der intuitiven Interpretation: Ähnliche Datenpunkte werden räumlich nah beieinander positioniert. Zudem eignet sich die Methode für verschiedene Distanzmetriken. Als Nachteile sind die hohe Rechenzeit bei großen Datensätzen sowie die Sensitivität gegenüber Ausreißern zu nennen (IU Internationale Hochschule, 2025, S. 89–94). Die Anwendung erfolgte auf den PCA-vorverarbeiteten Daten mit 50 Komponenten. Der normalisierte Stress-Wert diente als Gütekriterium für die Qualität der Projektion.

LLE als dritte Methode zielt darauf ab, lokale Nachbarschaftsstrukturen zu erhalten, indem jeder Datenpunkt als lineare Kombination seiner Nachbarn rekonstruiert wird. Der Vorteil liegt in der Fähigkeit, nicht-lineare Mannigfaltigkeiten zu entfalten und komplexe Strukturen sichtbar zu machen. Nachteilig ist die Sensitivität gegenüber der Wahl der Nachbaranzahl sowie die Anfälligkeit bei verrauschten Daten oder ungleichmäßig verteilten Stichproben (IU Internationale Hochschule, 2025, S. 98–102). Die Methode wurde mit 15 Nachbarn auf den PCA-reduzierten Daten angewandt, wobei der Rekonstruktionsfehler als Qualitätsmaß diente.

t-SNE eignet sich besonders für die Visualisierung hochdimensionaler Daten, da die Methode lokale Strukturen betont und Cluster visuell hervorhebt. Der Vorteil liegt in der oft beeindruckenden visuellen Trennung von Gruppen. Nachteile sind die fehlende Reproduzierbarkeit ohne festen Random-Seed, die hohe Rechenzeit sowie die Tatsache, dass globale Distanzen nicht erhalten bleiben und die Ergebnisse stark vom Perplexity-Parameter abhängen (Wattenberg et al., 2016). Die Anwendung erfolgte mit einer Perplexity von 30 auf den PCA-vorverarbeiteten Daten.

UMAP als modernste der eingesetzten Methoden kombiniert die Vorteile von t-SNE mit besserer Skalierbarkeit und Erhaltung globaler Strukturen. Die Methode basiert auf topologischen Konzepten und bietet schnellere Berechnungszeiten bei vergleichbarer Visualisierungsqualität. Nachteilig ist die Abhängigkeit von mehreren Hyperparametern sowie die geringere theoretische Fundierung im Vergleich zu klassischen Methoden (McInnes et al., 2018). Die Parameter wurden auf 15 Nachbarn und eine minimale

Distanz von 0,1 gesetzt.

Der Vergleich aller fünf Methoden zeigt deutliche Unterschiede in der Strukturerhaltung. PCA, t-SNE und UMAP liefern visuell gut separierte Cluster mit vier bis fünf erkennbaren Gruppen. MDS zeigt eine eher zusammenhängende Punktwolke ohne klare Trennung. LLE weist problematisches Verhalten auf: Fast alle Datenpunkte werden zusammengepresst, was auf eine ungeeignete Parameterwahl oder Datenstruktur hindeutet. Die reduzierten Datensätze aller Methoden wurden für das nachfolgende Clustering gespeichert, um die beste Kombination aus Dimensionsreduktion und Clustering-Algorithmus zu ermitteln.

2.3.3 Clustering

Für das Clustering wurden die reduzierten 2D-Datensätze aller fünf Dimensionsreduktionsmethoden geladen. Zur Bestimmung der optimalen Clusteranzahl kamen die Elbow-Methode und die Silhouette-Analyse zum Einsatz, wobei k-Werte von 2 bis 10 evaluiert wurden. Die Elbow-Methode analysiert den Abfall der Inertia (Within-Cluster Sum of Squares), während der Silhouette-Score die Qualität der Clusterzuordnung misst (IU Internationale Hochschule, 2025, S. 45–52). Basierend auf beiden Metriken wurde k=3 als optimale Clusteranzahl gewählt.

k-Means wurde als erster Clustering-Algorithmus auf allen fünf reduzierten Datensätzen angewandt. Der Algorithmus ordnet jeden Datenpunkt iterativ dem nächsten Clusterzentrum zu und verschiebt die Zentren, bis die Abstände innerhalb der Cluster minimal sind. Vorteile sind die einfache Implementierung, schnelle Konvergenz und gute Skalierbarkeit. Nachteilig wirken sich die Sensitivität gegenüber der Initialisierung, die Annahme sphärischer Cluster sowie die Notwendigkeit einer vordefinierten Clusteranzahl aus (IU Internationale Hochschule, 2025, S. 37–52).

Literatur

Behrens, S. (2025). *Sentimentanalyse*. Verfügbar 10. Januar 2026 unter https://github.com/svenb23/Sentimentanalyse

IU Internationale Hochschule. (2025). *Maschinelles Lernen – Unsupervised Learning und Feature Engineering* [DLBDSMLUSL01_D].

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Open Sourcing Mental Illness. (2016). *Mental Health in Tech Survey 2016*. Verfügbar 9. Januar 2026 unter https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to Use t-SNE Effectively. *Distill*. https://doi.org/10.23915/distill.00002

World Health Organization. (2025). *Over a billion people living with mental health conditions – services require urgent scale-up*. Verfügbar 9. Januar 2026 unter https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up

Verzeichnis der Anhänge

Anhang

Tabelle 1: Verteilung der Datentypen nach Skalenniveau

Haupttyp	Untertyp	Anzahl
Kategorisch	Nominal	42
Kategorisch	Ordinal	7
Kategorisch	Binär	7
Kategorisch	Multi-Value	4
Text	Freitext	2
Numerisch	Verhältnis	1

Tabelle 2: Spalten mit fehlenden Werten (Auszug)

Anzahl	Anteil	Spalte
1286	89,9%	Revealed to client impacted negatively
1226	85,7%	Work time affected percentage
1167	81,6%	Primary role related to tech/IT
1143	79,9%	Know local resources for help
1143	79,9%	Reveal diagnosis to clients
1143	79,9%	Revealed to coworker impacted negatively
1143	79,9%	Productivity affected by mental health
1109	77,6%	Conditions believed to have
863	60,3%	Diagnosed conditions
775	54,2%	Observations made less likely to reveal
721	50,4%	Professional diagnosed conditions
592	41,4%	US state (live)
581	40,6%	US state (work)
420	29,4%	Know employer coverage options
287	20,1%	Comfortable discussing with coworkers