Projekt: Maschinelles Lernen - Unsupervised Learning und Feature Engineering DLBDSMLUSL01_D

Fallstudie

Studiengang: Angewandte Künstliche Intelligenz

Sven Behrens

Matrikelnummer: 42303511

Prof. Dr. Christian Müller-Kett

9. Januar 2026

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

GMM Gaussian Mixture Models

LLE Locally Linear Embedding

MDS Multidimensional Scaling

OSMI Open Sourcing Mental Illness

PCA Principal Component Analysis

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.

Literatur

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

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