



## Projektbericht

Studiengang  
”Angewandte Künstliche Intelligenz”

**Computer Vision**  
**DLBAIPCV01\_D**

Dawid Jedlinski  
Matrikelnummer: IU14113900  
dawid.jedlinski@iu-study.org

Tutor: Ahmet Nasri  
Abgabedatum: 12.10.2025

# Inhaltsverzeichnis

<b>I. Abbildungsverzeichnis</b>	<b>IV</b>
<b>II. Tabellenverzeichnis</b>	<b>V</b>
<b>III. Abbreviations</b>	<b>VI</b>
<b>1. Einleitung</b>	<b>1</b>
1.1. Problemstellung . . . . .	1
1.2. Zielsetzung . . . . .	1
1.3. Vorgehensweise . . . . .	1
<b>2. Datenbeschreibung und EDA</b>	<b>2</b>
2.1. Herkunft und Struktur des Datensatzes . . . . .	2
2.2. Erste deskriptive Analysen??? . . . . .	2
2.3. Explorative Visualisierungen . . . . .	2
<b>3. Datenvorverarbeitung</b>	<b>3</b>
3.1. Umgang mit fehlenden Werten . . . . .	3
3.2. Bereinigung unstandardisierter Texteingaben??? . . . . .	3
3.3. Kodierung und Transformation der Merkmale??? . . . . .	3
<b>4. Feature Engineering</b>	<b>4</b>
4.1. Feature Selection . . . . .	4
4.2. Feature Generation . . . . .	4
<b>5. Dimensionsreduktion</b>	<b>5</b>
5.1. Methoden der Dimensionsreduktion . . . . .	5
5.2. Ergebnisse und Visualisierung . . . . .	5
<b>6. Clustering</b>	<b>6</b>
6.1. Auswahl geeigneter Methoden . . . . .	6
6.2. Bestimmung der Clusteranzahl . . . . .	6
6.3. Ergebnisse . . . . .	6
6.4. Übertragung auf HR-Kontext . . . . .	6
<b>7. Diskussion</b>	<b>7</b>
7.1. Bewertung des Vorgehens . . . . .	7
7.2. Grenzen der Analyse . . . . .	7
<b>8. Schluss</b>	<b>8</b>
8.1. Zentrale Erkenntnisse . . . . .	8

8.2. Ableitungen konkreter Maßnahmen für HR . . . . .	8
8.3. Ausblick . . . . .	8

## **I. Abbildungsverzeichnis**

## **II. Tabellenverzeichnis**

### **III. Abbreviations**

<b>AFL</b>	American Fuzzy Lop
<b>API</b>	Application Programming Interface
<b>BIOS</b>	Basic Input/Output System
<b>Brick</b>	Binary Run-time Integer Based Vulnerability Checker
<b>CaaS</b>	Container as a Service
<b>CAB</b>	Change Advisory Board
<b>CE</b>	Community Edition
<b>CI</b>	Continuous Integration
<b>CLI</b>	Command Line Interface
<b>CNCF</b>	Cloud Native Computing Foundation
<b>CRED</b>	C Range Error Detector
<b>Dev</b>	Development, the development team

# **1. Einleitung**

blablabla

## **1.1. Problemstellung**

- Bedeutung psychischer Gesundheit in technologiebezogenen Berufen
- Beschreibung des unternehmensinternen Präventivprogramms
- Herausforderungen: hohe Dimensionalität, fehlende Werte, unstrukturierter Text

## **1.2. Zielsetzung**

- Aufbereitung der Daten für bessere Interpretierbarkeit
- Reduktion der Komplexität durch Dimensionsreduktion
- Clustering zur Identifikation relevanter Gruppen
- Visualisierungen zur Unterstützung der HR-Entscheidungen
- Ableitung potenzieller Ansatzpunkte für das Präventionsprogramm

## **1.3. Vorgehensweise**

Übersicht über die Arbeitsschritte:

EDA → Datenbereinigung → Feature Engineering → Dimensionsreduktion → Clustering → Interpretation

## **2. Datenbeschreibung und EDA**

### **2.1. Herkunft und Struktur des Datensatzes**

- Quelle (z. B. Kaggle OSMI Mental Health in Tech 2016)
- Stichprobe, Anzahl der Merkmale, Datentypen
- Besonderheiten: Freitextfelder, kategoriale Felder, sensible Daten

### **2.2. Erste deskriptive Analysen???**

- Verteilungen wichtiger Merkmale
- Häufigkeiten, zentrale Tendenzen
- Identifikation möglicher Probleme: Outlier, Inkonsistenzen

### **2.3. Explorative Visualisierungen**

- Histogramme, Barplots, Boxplots
- Korrelationen / Heatmaps
- Erste Hypothesen über Muster im Datensatz

## **3. Datenvorverarbeitung**

### **3.1. Umgang mit fehlenden Werten**

- Identifikation der fehlenden Werte
- Strategien (z. B. Dropping, Imputation, Domain-Knowledge)
- Begründung der gewählten Methode

### **3.2. Bereinigung unstandardisierter Texteingaben???**

- Vereinheitlichung von Kategorien
- Lowercasing, Mapping, Domain-basierte Zusammenführung
- Umgang mit Freitext-Antworten

### **3.3. Kodierung und Transformation der Merkmale???**

- One-Hot-Encoding, Ordinal Encoding, ggf. Target-Encoding
- Skalierung (Transformation)
- Herausforderungen bei hochkardinalen Features

## **4. Feature Engineering**

### **4.1. Feature Selection**

- Variance Threshold
- Korrelationen / Redundanz
- Relevanzbasierte Auswahl (Mutual Information)

### **4.2. Feature Generation**

- Erstellen neuer Merkmale aus bestehenden Variablen
- Beispiele: Stress-Score, Support-Index, Arbeitsumfeld-Indikatoren
- Nutzen für Modellverständlichkeit und Clustering

## **5. Dimensionsreduktion**

Warum Dimensionsreduktion?

Vorgehensweise

### **5.1. Methoden der Dimensionsreduktion**

- PCA (linear)
- MDS, LLE (nichtlinear)
- Vergleich und Begründung der Auswahl

### **5.2. Ergebnisse und Visualisierung**

- Erklärte Varianz (PCA)
- 2D/3D-Darstellungen
- Herausgearbeitete Muster und Trends

## **6. Clustering**

### **6.1. Auswahl geeigneter Methoden**

- K-Means
- Agglomeratives Clustering
- DBSCAN/HDBSCAN für komplexe Strukturen
- Begründung der Auswahl

### **6.2. Bestimmung der Clusteranzahl**

- Elbow-Methode
- Silhouette Score
- Weitere Metriken

### **6.3. Ergebnisse**

- Visualisierungen der Cluster (PCA/UMAP Scatterplots)
- Profiling: Beschreibung der typischen Merkmale jedes Clusters
- Identifikation gefährdeter Gruppen und Muster

### **6.4. Übertragung auf HR-Kontext**

- Welcher Cluster ist besonders belastet?
- Welche Kombinationen von Faktoren treten gehäuft auf?
- Welche Gruppen könnten gezielte Unterstützung benötigen?

## **7. Diskussion**

### **7.1. Bewertung des Vorgehens**

- Was hat gut funktioniert?
- Was hat schlecht funktioniert?
- Welche Alternativen wären möglich?

### **7.2. Grenzen der Analyse**

- Qualität der Umfragedaten
- Generalisierbarkeit
- Nicht berücksichtigte Faktoren

## **8. Schluss**

### **8.1. Zentrale Erkenntnisse**

- Welche Cluster wurden gefunden?
- Was sind deren Hauptmerkmale?
- Welche Muster sind besonders problematisch?

### **8.2. Ableitungen konkreter Maßnahmen für HR**

- Zielgruppenspezifische Interventionen
- Programme zur psychischen Entlastung
- Verbesserungen von Arbeitsbedingungen
- Informations- und Unterstützungsangebote

### **8.3. Ausblick**

- Nutzung weiterer Datenquellen
- Kontinuierliches Monitoring
- Potenzial für zukünftige ML-Modelle

### **Literaturverzeichnis**

#### **Anhang** - Visualisierungen

- Feature-Listen
- Clustering-Parameter

### **LINK ZU GITHUB!**