INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Projektbericht

Studiengang
”Angewandte Künstliche Intelligenz”

Computer Vision
DLBAIPCV01_D

Dawid Jedlinski

Matrikelnummer: IU14113900

dawid.jedlinski@iu-study.org

Tutor: Ahmet Nasri

Abgabedarum: 12.10.2025

Inhaltsverzeichnis

I. Abbildungsverzeichnis

II. Tabellenverzeichnis

III. Abbreviations

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

Im Rahmen der Exploratory Data Analysis wurde der Datensatz auf Struktur, Verteilungen, fehlende Werte und potenzielle Inkonsistenzen untersucht. Dabei wurden zentrale Merkmale analysiert und erste Muster identifiziert, die Hinweise auf relevante Einflussfaktoren psychischer Belastung liefern. Die Ergebnisse der EDA bilden die Grundlage für die anschließende Vorverarbeitung und das Feature Engineering.

2.1. Herkunft und Struktur des Datensatzes

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

- Also Quelle ist https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016?resource=download
- Die OSMI Mental Health in Tech Survey 2016 ist eine internationale Umfrage mit über 1400 Teilnehmern aus dem IT- und Tech-Bereich. Ziel der Umfrage ist es, Einstellungen gegenüber psychischer Gesundheit am Arbeitsplatz zu erfassen und die Häufigkeit psychischer Erkrankungen unter Beschäftigten in der Tech-Branche zu untersuchen. Die gesammelten Daten werden vom Open Sourcing Mental Illness (OSMI) Team genutzt, um das Bewusstsein für psychische Gesundheit zu stärken und die Arbeitsbedingungen für Betroffene in der IT zu verbessern.
- 1433 Zeilen also Teilnehmer und 63 Spalten also Fragen.
- Datentypen sind int64, float64 und hauptsächlich object also eine Texteingabe.
- Fehlende Werte gibts i.d.R. viele (siehe Heatmap). Zwischen Fragen 0 bis 36 gibts immer wieder Leute die nichts eingetragen haben. Die meisten fehlende Werte liegen zwischen Frage 16 und 24. Das sind Fragen wie:
  - ”Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?
  - If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?
  - Do you believe your productivity is ever affected by a mental health issue?
  Also konkrete und stark private Fragen, die jedoch am meisten zum Thema beitragen

2.2. Erste deskriptive Analysen???

- Verteilungen wichtiger Merkmale (Gender, Age, Wohnland/Arbeitsland)
- Identifikation möglicher Probleme: Outlier, Inkonistenzen

- es wurden basisdaten analysiert wie GENDER, AGE, WOHNLAND und Arbeitsland

- Da GENDER keine vordefinierte Antworten hatte, hat jeder Befragte seine eigene Antwort geschrieben, was dazu führt dass danach eingie Gruppen zusammengeführt werden müssen (wie Female, female, f) und andere komplett entfernt (z.B. Dude, mail)
- beim AGE gibts Ausreißer über 100 und 300 Jahre analysiert
- Arbeitsland und Wohnland sehen gut aus, hier zeigt sich dass der Großteil aus USA und UK kommt
- Die Befragten arbeiten hauptsächlich in dem Land in dem sie wohnen (1407:26)

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

OFFENE FRAGEN LÖSCHEN WIE ”Why or why not?” - 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!