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. [9]
- Datentypen sind int64, float64 und hauptsächlich object also eine Texteingabe. [7]
- 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 [43] 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 [37], AGE[38], WOHNLAND[40] und Arbeitsland[41]

- 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 [39]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 (siehe Heatmap)
- Strategien (z. B. Dropping, Imputation, Domain-Knowledge)
- Begründung der gewählten Methode

- zuerst werden alle offenen Fragen gelöscht die schwer von KI zu interpretieren sind für eine Clustering Aufgabe (z.B. Why or why not?) und Fragen die auf vorherige Antwort bezogen sind also (What US state do you work in/live in?) [114]
- bezüglich fehlenden Werten werden zuerst Befragte gesucht die von allen Fragen mehr als 40% unbeantwortet haben. Diese entfernen! [115]
- jetzt die Fragen die mehr als 40% missing ratio haben werden gelöscht [117]
- Befragte die weniger als 25%fehlende Werte haben, werden durch Imputation ergänzt. Hier werden mehrere Antwortarten und Imputationen unterschieden: [119, 122]
  - kategorial (ja/nein/idk) – idk selbst wählen
  - ordinal (gut/mittel/schlecht) – median(), zuerst in zahlen kodieren, dann Median berechnen und dann entkodieren
  - numerisch (z.B. Alter) – median()
  - multilabel (z.B. Job-Rollen) – ”UNKNOWN”
  - binary (0/1. ja/nein) – mode()

  Es wurden insgesamt 1014 und davon wurden nur die berücksichtigt die tatsächlich fehlende werte haben. Von den Befragten (missing ratio ¡ 0.25) wurden dann 3 Fragen gefunden die fehlende Werte hatten
- Frage 4: kategorial: aus no/notsure/yes, wird automatisch der wert notsure zugewiesen
- Frage 32: auch kategorial: aus no/maybenotsure/yesiobserved/yesiexperienced wird automatisch maybe/notsure zugewiesen
- Frage 41: sollten eigentlich drei vordefinierte Werte sein - male, female, others. Dann fehlen nur 3 Werte und es werden einfach zu Others hinzugewiesen. Der Umgang mit verschiedenen Kategorien (also Vereinheitlichen) wird im nächste Kapitel beschrieben.

3.2. Ausreißer und Werte vereinheitlichen

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

- jeder hat freien text geschrieben und es sind antworten gekommen wie ”f”, ”cis man” ”none of your business”, deswegen bevor die fehlenden werte imputiert werden, müssen die antworten vereinheitlicht werden (durch mapping). Es wurden keywords vordefiniert z.b. bei man (male, m, man, ...). Falls solche in der Antwort vorkommen, wird es zu Man gemappt, genauso mit Female. Alles andere wird zu Others zugewiesen [119]
- beim age werden ausreißer und unseriöse werte. Allgemein werden also Befragte mit Alter ¡17 und ¿67 gelöscht [124]
- Einzelwerte von AGE wurden in Gruppen zusammengefasst (17-25, 26-35, 36-45, 46-55, 56-67) [124]
- länder (WOHNLAND und ARBEITSLAND) werden zu 10 am häufigsten vorkommenden Ländern zusammengefasst, der Rest zu Others [124]
- bei Jobrollen gibts sehr viele Angaben (auch mehrere Rollen pro Befragten) Diese werden zu übergeordneten Gruppen zugewiesen (z.B. Backend/Frontend Developer einfach zu Developer). es werden 8 Hauptgruppen unterschieden; Management/Lead, Developer, DevOps, Product Design, Data & Analytics, HR/Admin, Community, Other. Da viele Personen mehrere ”Rollen” haben wird noch eine Priorität bei der Auswahl gewählt: Lead=1, Developer=2, usw. bis Community=7 und Other=99

3.3. Kodierung und Transformation der Merkmale???

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

- binäre Daten - OneHot, kategorial nominale - OneHot, kategoriale ordinale - OrdinalEncoder
- für binäre/nominale daten wurden zuerst manuell alle antworten analysiert und nur die spalten gewählt die diese art von antworten haben, danach transformiert fitten.
- ordinale sind alle restlichen (werte wie ”I dont know werden immer ganz hinten liegen, modelle können es dann als ’separate kategorie berachten’. es wird nicht gelöscht, da viele personen solche antwort gewählt haben und diese kann sehr wertvoll sein)
- für onehot wurde neues dataframe erstellt, für die ordinale wurde data_clean überschrieben.
- Mit pd.concat() wurden die beiden kombiniert zu einem neuen dataframe data_encoded, welches nur aus numerischen werten besteht
- ordinalencodierte und onehotencodierte skalieren wegen Einheitlichkeit für abstandbasierte ML-Modelle. Dies erfolgt mit StandardScaler aus sklearn.preprocessing. Es wird nicht mit MinMax gemacht, da ...
- der vorverarbeitete datensatz wurde als neue datei gespeichert data_preprocessed.csv

4. Feature Engineering

4.1. Feature Selection

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

- für unsupervised learning eignen sich am besten der Variance Threshold und die Korrelationsmatrix
- zuerst variance threshol mit einem schwellenwert von 0.01. Es wurden zwei Fragen gefunden die binär waren und nur eine Antwort haben
  - Are you self-employed?_0 mit allen 0 bzw False Werten(_1 gibts nicht, da alle NEIN beantwortet haben)
  - Do you have previous employers?_1 mit allen 0 bzw False (_0 gibts nicht, da alle auch NEIN beantwortet haben)
  also kann man beide löschen, da sie immer die gleichen werte haben
- es wurden bei korrelationsmatrix mehrere fragen automatisch gewählt.
- - Is your employer primarily a tech company/organization?_0.0
  - Is your employer primarily a tech company/organization?_1.0: 1.00
  Es ist eine binäre Frage die one-hot encodiert ist, und somit wenn ich eine davon lösche, verliere ich keine Informationen von den anderen
  das gleiche gilt für
  - Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?_No
  - Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?_Yes: 1.00
  und noch diese
  - Have you been diagnosed with a mental health condition by a medical professional?_No
  - Have you been diagnosed with a mental health condition by a medical professional?_Yes: 1.00
  und diese - Have you ever sought treatment for a mental health issue from a mental health professional?_0
  - Have you ever sought treatment for a mental health issue from a mental health professional?_1: 1.00

- bezüglich Wohnland und Arbeitsland wird die jede Spalte vom Wohnland entfernt, da die kleinste Korrelation einen wert von 0.96 beträgt, was sehr hoch Inkonistenzen
- AUF FEATURE GENERATION WURDE VERZICHTET, WEIL...

4.2. Methoden der Dimensionsreduktion

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

4.2.1. ANSATZ A: PCA

- WARUM DIMENSIONSREDUKTION?
- PCA wird verwendet, weil sie auf varianz basiert und alle werte sind bereits encodiert und standardisiert auf mittelwert 0 und varianz 1
- die Features wurden linear skaliert mit StandardScaler, daher sind die für lineare Analyse gut vorbereitet (PCA ist linear)
- MDS nicht, weil distanzbasiert und ich muss zuerst ein Distanzmaß definieren
- LLE nicht, weil setzt Mannigfaltigkeiten voraus, das trifft bei Fragebogendaten fast nie zu, benötigt viele Datenpunkte item zuerst wird ein Explained Variance Plot dargestellt um die beste Anazahl von Hauptkomponenten zu finden
- der Ellenbogen Punkt liegt bei 10, jedoch es ist unter 85% Varianz, diese befindet sich erst bei 64 Hauptkomponenten. 10 enthält nur 32% der Varianz. Es wird also K=64 gewählt.
- DIAGRAMME ZEIGEN
- PROBLEM: sehr viele Hauptkomponenten, schwierig interpretierbar, UND wenn Anzahl Cluster Berechnung mit BIC ist ein sehr hoher Wert (kleinster bei k=2 220000) und SilhouettenScore sehr niedrig (größter bei k=2 0.044)
- aus diesen Gründen wurde Ansatz B gewählt

4.2.2. ANSATZ B: manuelle Feature Transformation

- hierfür wurden Fragen die zu gleichen Themen gehören zusammengefasst, Antworten gewichtet, in neuer Feature gespeichert und danach gelöscht
- NEUE Features:
- employer_support_score - höher je mehr Unterstützung der Arbeitgeber bietet
- prev_employer_support_score - höher je mehr Unterstützung der ehemalige Arbeitgeber geboten hat
- openness_score - wie offen der Befragte bei einem neuen Arbeitgeber wäre
- perceived_stigma_score - höher, wenn der Befragte denkt, dass MH schadet (Karriere/Team)
- mh_status_score - misst wie stark der Befragte mental gesund ist

- jede Antwort wird entsprechend bewertet, werte die gegen den Score sind MINUS und Werte für den Score sind PLUS. Danach wird die Summe aller Fragen berechnet, die zu dem Score gehören und der Durschnitt berechnet (danach mit StandardScaler skaliert)
- insgesamt 106 Fragen zu 5 Features zusammengefasst und danach Fragen gelöscht
- VORTEILE: jetzt sind es insgesamt 35 Features, also weniger als 64 bei PCA, die Features sind gut interpretierbar und außerdem haben die besseren BIC (k=2 35000) und silhouetten (k=2 0.29)
- bic sinkt mit anzahl cluster (also laut bic k=12 besser als k=2), beim silhouetten ist umgekehrt. Es werden k=3 gewählt (also eher silhouetten betrachten), da es angibt wie gut getrennt und ziel der aufgabe ist gute interpretierbarkeit. Es wird k=3 gewählt, da laut bic je mehr desto besser und bei silhouetten ist k=2 (0.29) und k=3 (0.28) sehr ähnlich, k=4 (0.13) sinkt sehr stark, deswegen ist k=3 perfekt

5. Clustering

5.1. Auswahl geeigneter Methoden

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

- berücksichtigt wurden folgende Algorithmen: k-Means, GMM, DBSCAN und hierarchisch Agglomerativ
- VERGLEICH QUELLEN:
- https://learninglabb.com/clustering-algorithms-in-machine-learning/
- https://krishnapullak.medium.com/guide-to-clustering-algorithms-strengths-weaknesses-and-evaluation-5285a75ea902
- https://www.geeksforgeeks.org/data-science/choosing-the-right-clustering-algorithm-for-your-dataset/
- es wurde GMM gewählt, da KMEANS dafür zu ”fest” ist, andere waren für die Datensätze nicht geeignet.
-

5.2. Bestimmung der Clusteranzahl

- Elbow-Methode
- Silhouette Score
- Weitere Metriken

-

5.3. Ergebnisse

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

5.4. Übertragung auf HR-Kontext

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

6. Diskussion

6.1. Bewertung des Vorgehens

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

6.2. Grenzen der Analyse

- Qualität der Umfragedaten
- Generalisierbarkeit
- Nicht berücksichtigte Faktoren

7. Schluss

7.1. Zentrale Erkenntnisse

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

7.2. Ableitungen konkreter Maßnahmen für HR

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

7.3. Ausblick

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

Literaturverzeichnis

Anhang - Visualisierungen
- Feature-Listen
- Clustering-Parameter

LINK ZU GITHUB!