Mental Health in Technology-related Jobs

NEGIN HEZARJARIBI

Machine Learning – Unsupervised Learning and Feature Engineering (DLBDSMLUSL01)

Studies: Bachelor in Applied Artificial Intelligence

IU INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES (IU) | 31 DECEMBER 2024

Table of Contents

List of Figures and Tables ..... *II*

Introduction **1**

1. Methodology **1**

1.1 Data Overview 1
1.2 Data Cleaning and Preprocessing 2
1.3 Column-Specific Data Cleaning and Feature Engineering 3
1.4 Feature Selection 4
1.5 Standard Numerical Data 4
1.6 Dimensionality Reduction 5
1.7 Clustering and Insights 5
1.8 Challenges in the Clustering Process 8

2. Findings and Recommendations **9**

Conclusion **10**

Tools and Libraries Used *11*

References *13*

List of Figures and Tables

FIGURE 1:VISUALIZING UMAP DATA AFTER FEATURE SELECTION IN 3D WITH GMM CLUSTERING 5
FIGURE 2: HEATMAP OF INTENSITY OF MEAN VALUES OF FEATURES ACROSS CLUSTERS 6
FIGURE 3: BAR CHART OF STANDARD DEVIATION OF SIGNIFICANT FEATURES ACROSS CLUSTERS 7
FIGURE 4: BAR CHART OF COMPARISON OF SIGNIFICANT FEATURES ACROSS CLUSTERS 8

Introduction

Mental health plays a crucial role in our overall well-being, particularly in high-pressure fields like technology. The World Health Organization (2022) emphasizes that good mental health allows individuals to tackle challenges, utilize their skills, work effectively, and make positive contributions to their communities. This highlights its importance for both personal and professional success.

In light of these challenges, our company's HR department is committed to addressing the mental health concerns of employees. As a data scientist, I was assigned the task of analyzing a mental health survey conducted among tech professionals. The dataset is intricate, featuring numerous variables, missing data, and unstructured text. My objective is to streamline this data, categorize participants based on their responses, and deliver actionable insights through visualizations to assist HR in enhancing employee well-being and productivity.

1. Methodology

This section describes the approaches taken to tackle the challenges presented by the dataset, such as missing data, high dimensionality, and mixed data types. The workflow included feature selection, dimensionality reduction, and data normalization to prepare the dataset. Clustering methods and evaluation metrics were employed to ensure the results were both accurate and interpretable.

1.1 Data Overview

The Mental Health in Tech Survey 2016 (OSMI, 2016) dataset, obtained from Kaggle, comprises responses from 1,433 participants regarding their attitudes and challenges related to mental health in the tech industry. A detailed implementation, including the Python libraries utilized, can be found in the GitHub notebook (https://github.com/NeginHz/DLBDSMLUSL01).A "List of Tools and Libraries Used" is also included in the case study appendix.

1.1.1 Dataset Characteristics and Challenges

The dataset includes 63 columns that feature both qualitative and numerical data:

Qualitative Data (56 columns):

Nominal Variables: Categories without a specific order, such as Gender.

Ordinal Variables: Ordered categories, like the ease of requesting medical leave for mental health reasons (e.g., "Somewhat easy," "Very easy")

Numerical Data (7 columns):

Binary responses (e.g., 1 for Yes, 0 for No) and continuous values like Age.

This dataset presents challenges like high dimensionality, missing data, and a mix of data types, necessitating techniques such as dimensionality reduction, imputation, and normalization for effective analysis. It's crucial to simplify the data while preserving key features to derive actionable insights.

1.2 Data Cleaning and Preprocessing

1.2.1 Column Removal Based on Missing Data

Twelve columns were eliminated due to significant missing data and low relevance. Nine columns had over 79% missing data, rendering imputation unreliable. For instance, the column “Is your primary role within your company related to tech/IT?” (81.6% missing) was removed because a similar column, “Which of the following best describes your work position?”, had complete data, ensuring no gaps in the dataset.

Furthermore, the columns related to mental health discussions in the workplace (“Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?”, “Why or why not?”, and “Why or why not?.1”) were also discarded. Although they had 44%, 23%, and 21% missing data respectively, they were excluded due to difficulties in meaningful imputation and their minimal impact on the overall analysis.

1.2.2 Addressing Missing Values

Effectively handling missing data requires understanding its causes and applying appropriate strategies (Rubin, 1976). In this analysis, missing values in employer-related columns were linked to the “Are you self-employed?” question, indicating the data were Missing Not At Random (MNAR) (Little & Rubin, 2020). To address this, categorical columns were filled with “Not Applicable” for missing employer data, and numerical columns, like “Is your employer primarily a tech company?”, were replaced with -1 to differentiate from valid responses.

Missing data in previous employer columns was linked to the question “Do you have previous employers?”, with missing values appearing only when respondents had no prior employers. These were replaced with “Not Applicable” or -1 based on the column type, indicating the lack of previous employment.

1.3 Column-Specific Data Cleaning and Feature Engineering

After addressing most of the missing values, I chose to move forward with data cleaning and feature engineering for each column separately. This approach, highlighted in best practices for data preparation, ensures that the unique characteristics and needs of each variable are thoughtfully considered to improve overall data quality and model performance (Han et al., 2011).

1.3.1 Handling Missing Values for Specific Cases

For the column *"What is your gender?"*, three missing entries were categorized as *"Undisclosed"*, ensuring inclusivity and avoiding unwarranted assumptions. Similarly, the column *"Do you know the options for mental health care available under your employer-provided coverage?"* had missing values imputed with *'I am not sure'*, reflecting a neutral stance toward uncertainty.

In the column *"If so, what condition(s) were you diagnosed with?"*, missing values were filled based on responses in the column *"Have you been diagnosed with a mental health condition by a medical professional?"*. Respondents who answered *No* were labeled as *"No Diagnosis"*. For the remaining five missing values, the most common diagnosis, *"Mood Disorder (Depression, Bipolar Disorder, etc.),"* was used to preserve data consistency.

1.3.2 Feature Combination

To enhance the dataset, two related columns "*If yes, what condition(s) have you been diagnosed with?*" and "*If maybe, what condition(s) do you believe you have?*"—were combined into a new column, "*Type of currently mental health disorders,"* to provide a comprehensive representation of respondents' mental health conditions. For respondents who selected "Maybe" in "*Do you currently have a mental health disorder?*" and lacked a recorded diagnosis, their possible condition from the second column was used to fill the gap. This process improved data quality by reducing missing values, consolidating related information, and simplifying analysis, making the dataset more suitable for modeling.

1.3.3 Consolidating Rare Categories

To enhance clarity and reduce sparsity in categorical data, rare categories were consolidated into broader groups. The gender variable was consolidated into three primary categories: Male, Female, and Non-binary/Other. For country, the most frequent countries, namely the United States, United Kingdom, and Canada, were retained as separate categories, while all other countries were grouped under an "Other Country" category. In discussing mental health disorders, less common diagnoses were grouped into an "Other" category to keep the focus on more prevalent conditions.

1.3.4 Outlier Removal

In the column *“What is your age?”*, unrealistic entries, such as *3* and *323*, were removed. The age range was restricted to *15–99* based on logical and demographic considerations. After adjustment, the range narrowed to *15–74*, with a mean of *34.06*.

1.3.5 Data Transformation

To prepare the dataset for clustering algorithms, which rely on numerical input to calculate distances effectively, I transformed non-numerical columns into numerical ones using two main methods, Label Encoding and One-Hot Encoding based on the data type:

Label Encoding for Ordinal Data: Ordinal columns, like “How many employees does your company have?” and “Do you work remotely?”, were encoded to preserve their order. For example, "1-5" → 1, "6-25" → 2, and so on in “How many employees does your company have?” column.

One-Hot Encoding for Nominal Data: Nominal columns without a specific order, such as "Work position," "Mental health disorders," and "Gender," were one-hot encoded to avoid implying any hierarchy.

This method ensured the dataset was prepared for clustering algorithms while still being interpretable.

1.4 Feature Selection

To optimize the dataset for clustering, a filter-based method was used to remove highly correlated features (correlation coefficient > 0.85), addressing multicollinearity and redundancy. For example, strongly correlated columns like "What country do you live in?" and "What country do you work in?" resulted in the removal of one. A correlation matrix was utilized to keep relevant, independent features, reducing the risk of overemphasizing certain variables and enhancing clustering performance, as indicated by a higher Silhouette Score. This methodical improvement guaranteed more precise and meaningful clustering results.

1.5 Standard Numerical Data

Since algorithms like K-Means and GMM rely on distance metrics and are sensitive to scale differences, numerical data was standardized using StandardScaler to ensure fair feature representation during clustering. To avoid bias in clustering due to varying feature ranges, standardization is essential. For instance, features like Age (15–90) could dominate smaller range features (-1 to 5). Standardization balanced feature contributions, increased algorithm accuracy, and improved clustering quality (Han et al., 2011).

1.6 Dimensionality Reduction

To simplify high-dimensional data and enhance clustering, UMAP (Uniform Manifold Approximation and Projection) was chosen after evaluating PCA, t-SNE, and UMAP. UMAP was was selected due to its ability to maintain data structure, handle nonlinear patterns, and process large datasets efficiently. This method improved clustering accuracy and interpretability, as supported by McInnes, Healy, & Melville (2020). The GitHub repository contains implementation details, including how to install the library (pip install umap-learn).

1.7 Clustering and Insights

I tested several approaches, such as K-Means and Gaussian Mixture Models (GMM), to determine the best clustering technique. I chose GMM because of its flexibility in handling clusters of different sizes, shapes, and densities. Also GMM better models the data distribution, making it a more suitable choice for our data's complexity.

After testing various cluster numbers, Based on a reasonable Negative Log-Likelihood (6346.18) for model fit and the highest Silhouette Score (0.8427) for cluster separation (Rousseeuw, 1987), I concluded that k=3 was the most appropriate (Bishop, 2006). These metrics confirmed well-defined, compact, and distinct clusters, ensuring meaningful insights.

The clustering findings with UMAP reduced data are displayed in three dimensions in Figure 1, emphasising the distinct division and organisation of the clusters. This demonstrates how well GMM captures the patterns in the data and how appropriate it is for the study.

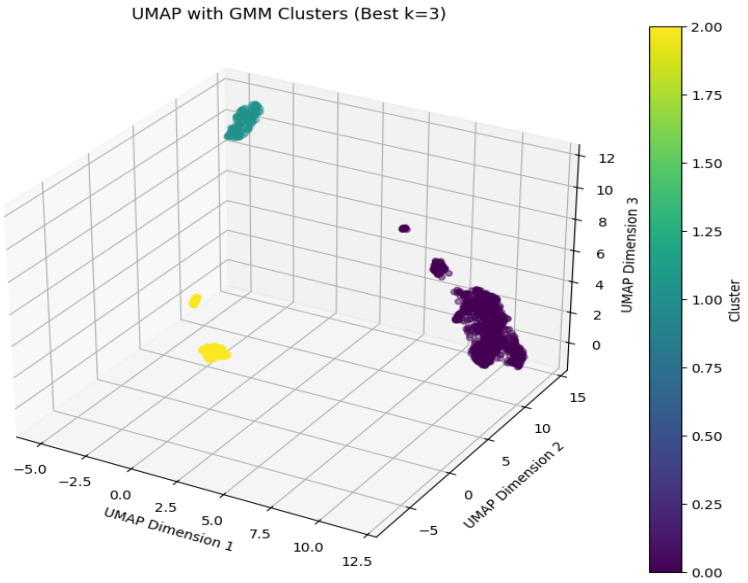

Figure 1:Visualizing UMAP data after feature selection in 3D with GMM clustering

1.7.1 Cluster Analysis

The Gaussian Mixture Model (GMM) was used to divide the dataset into three clusters, each of which has the following distribution: 70.77% of the data was in Cluster 0, 17.41% was in Cluster 1, and 11.82% was in Cluster 2. I performed a thorough examination of feature means and standard deviations in order to gain a deeper understanding of the characteristics of these clusters.

The analysis of feature means provided insights into the dominant traits defining each cluster(Wilkinson & Friendly, 2009). By examining the average values of features, I could identify which characteristics were most prevalent in each group. For example, the heatmap in Figure 2 shows the intensity of average feature values for each cluster; lower values are shown in blue, and greater values are shown in red. This approach is particularly useful for identifying patterns, such as whether a cluster has a higher mean age or specific tendencies in responses. Such patterns offer a clear and concise summary of the general attributes of each group, facilitating a more structured interpretation of cluster behaviors (Jain et al., 1999).

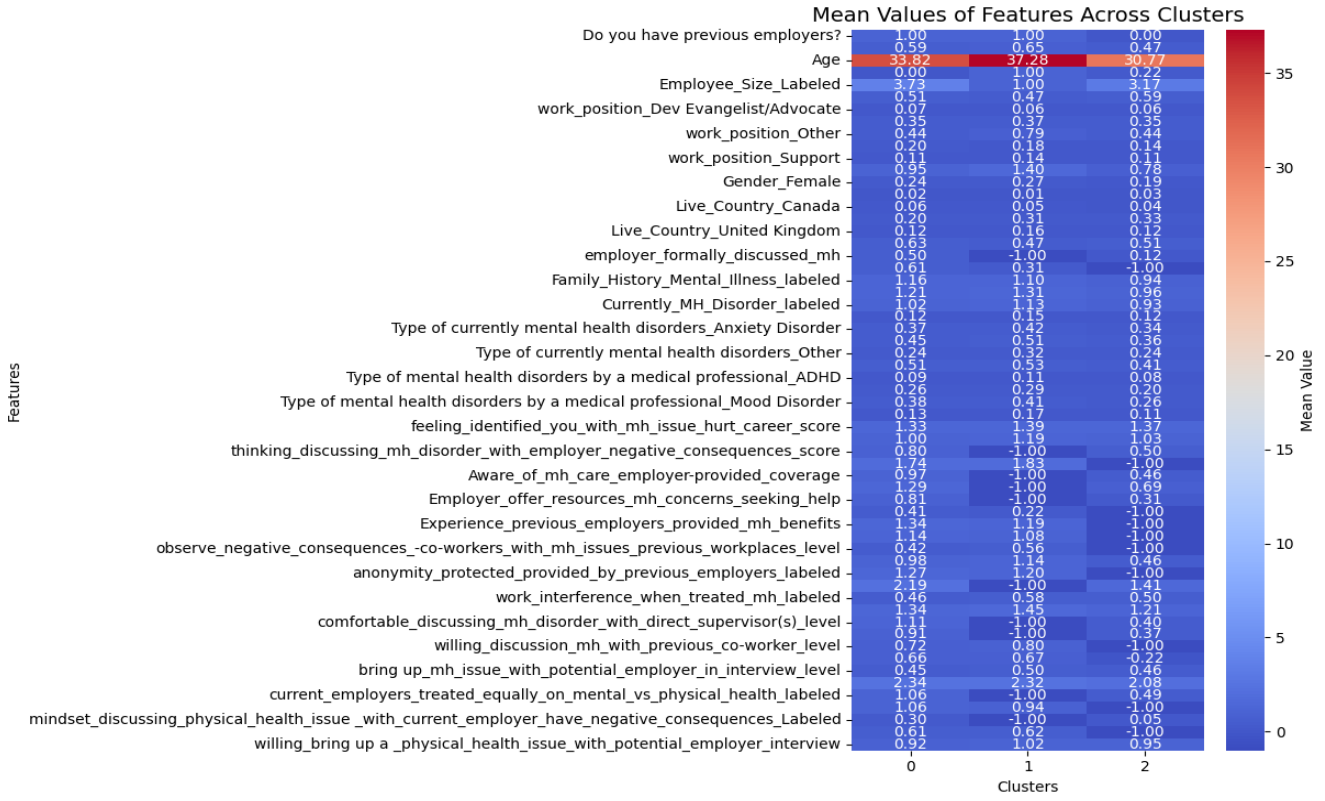

Figure 2: Heatmap of intensity of mean values of features across clusters

Additionally, analyzing the standard deviation of features within each cluster highlighted the variability in responses. This step helped uncover the degree of diversity or homogeneity in the data. For instance, Figure 3 illustrates the variation in significant features across clusters, showing how some features, like 'work_interference_when_NOT_treated_mh_labeled' showed higher variability, indicating diverse responses, while features like 'self-employed?' were more uniform, reflecting consistency within clusters.

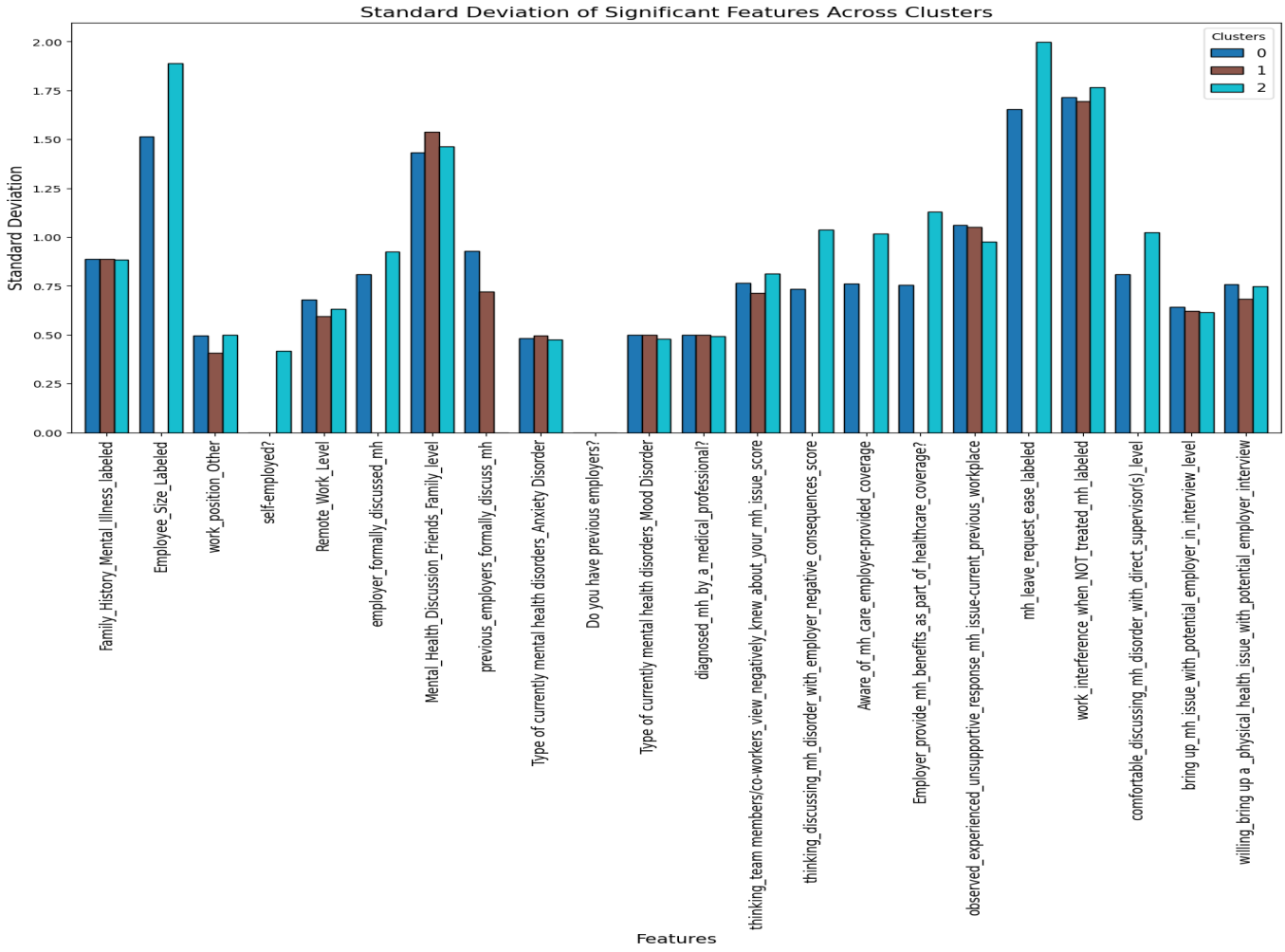

Figure 3: Bar Chart of Standard Deviation of Significant Features Across Clusters

Comparative bar charts showing the average values of significant features for each cluster were made in order to further visualise these findings (Figure 4: "Comparison of Significant Features Across Clusters"). Cluster 0 is represented in purple, Cluster 1 in green, and Cluster 2 in yellow. These visuals effectively highlight the differences in feature dominance across clusters, making it easier to interpret the results.

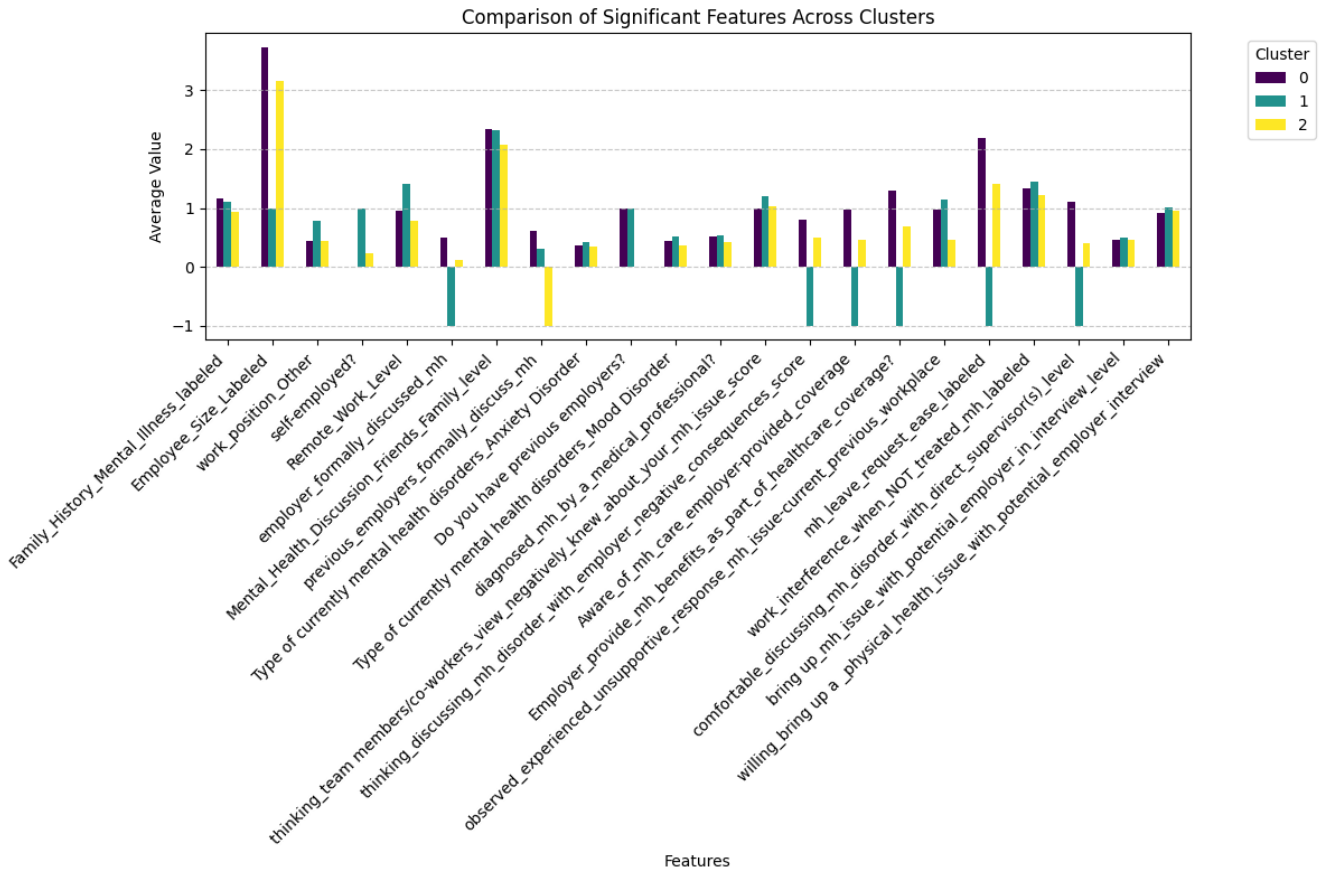

Figure 4: Bar Chart of Comparison of Significant Features Across Clusters

To clarify the interpretation of columns with an average value of -1,For example , I analyzed 'previous_employers_formally_discuss_mh' in Cluster 2 (yellow). The value of -1 reflects missing data imputed during preprocessing for participants without previous employers, as indicated by the related column 'Do you have previous employers?'. Figure 2 shows that Cluster 2's average value of 0 for this feature confirms the absence of prior employment. This preprocessing decision ensures the data remains interpretable and consistent with the feature engineering assumptions.

This analysis of means and standard deviations clarified cluster structures and diversity, highlighting key features and providing deeper insights into each group's unique traits for better clustering interpretation.

1.8 Challenges in the Clustering Process

The clustering process involved addressing several challenges, including high dimensionality, where a dataset with 63 columns required feature selection using the Filter Method and dimensionality reduction via UMAP for improved interpretability; missing data, which was managed through imputation and categorizing values like "Not Applicable" while ensuring the formation of meaningful groups; and data standardization, necessitating normalization, Label Encoding, and One-Hot

Encoding to accommodate mixed data types for compatibility with machine learning algorithms. Gaussian Mixture Models (GMM) were selected for their flexibility and ability to handle overlapping data, with the Silhouette Score and Elbow Method used to determine the optimal number of clusters. Future improvements could involve advanced imputation techniques, such as KNN or MICE, and deeper integration of additional datasets.

2. Findings and Recommendations

Through clustering analysis, I identified three distinct groups with unique characteristics. These insights form the foundation for customized recommendations to address the specific needs of each cluster effectively.

Cluster 0: The Experienced Professionals

With an average age of 33.8 years, this group consists of people in the middle of their careers. Every member has a 100% history of employer interactions and shows a high level of awareness regarding mental health coverage offered by their employers. They are most comfortable discussing mental health with friends and family, showing the highest scores in this aspect among all clusters. Members of this cluster generally work in larger, more structured organizations, with most identifying as Back-end Developers. Their concerns about discussing mental health in the workplace are relatively low.

Recommendations emphasize encouraging open dialogues through workshops or mental health support groups to build on individuals' comfort in structured environments, providing efficient tools and processes to enhance team collaboration, particularly for Back-end Developers, and implementing access to corporate psychologists and group mental health sessions in larger organizations to improve engagement.

Cluster 1: The Self-Employed and Cautious

This group has the greatest average age (37.2 years) and is fully self-employed (100%). Members express significant concerns about the negative repercussions of discussing mental health issues with employers and report a history of dissatisfaction with workplace support. They strongly enjoy working remotely and have greater incidence rates of mental health issues, especially anxiety and mood disorders. Their roles often fall outside traditional job categories.

Recommendations focus on improving access to mental health resources by providing remote counseling sessions and online services designed for independent workers, developing long-term support solutions such as stress and anxiety management programs via apps and self-help tools that

respect their need for autonomy, and creating focused campaigns to address common mental health issues like mood disorders and anxiety with easily.

Cluster 2: The Emerging Workforce

Younger people (average age: 30.7 years) without previous employer relationships are included in this group. They report moderate comfort levels with remote work and have average concerns about disclosing mental health issues. Their employers often provide limited mental health support, and their roles frequently include Back-end Development.

Recommendations include conducting workshops to reduce stigma and encourage a more open discussion about mental health challenges, offering career coaching and soft skills workshops to support individuals in navigating early-career challenges, and providing online mental health resources and training programs to address the lack of employer-supported benefits in smaller companies.

Comparative Analysis and Conclusion

The clusters highlight diverse mental health awareness, workplace experiences, and needs:

Cluster 0: Well-supported professionals benefiting from structured environments.

Cluster 1: Self-employed individuals facing higher mental health challenges and preferring autonomy.

Cluster 2: Early-career workers needing foundational support and resources.

Customized programs addressing these unique characteristics can significantly enhance employees' mental well-being, productivity, and job satisfaction. HR departments can develop focused initiatives to promote a more supportive and healthy work environment by putting these suggestions into practice.

Conclusion

This case study analyzed mental health patterns among technology professionals using clustering techniques to identify actionable insights for HR teams. By addressing challenges in data preprocessing and applying robust clustering methods, I developed meaningful group profiles to inform targeted interventions. These findings emphasize the importance of personalized mental health programs that address the unique needs of different workforce segments, contributing to a healthier and more supportive work environment.

To gain a deeper understanding of the process and access detailed steps, you can explore the full project on my GitHub repository (https://github.com/NeginHz/DLBDSMLUSL01).

Tools and Libraries Used

This case study utilized the Python programming language for data analysis and interpretation, utilizing its rich ecosystem of libraries to support each phase of the project. The analysis was conducted using Google Colab, an interactive and efficient platform for writing and executing code in a collaborative environment.

Python and its extensive libraries were instrumental in handling data preprocessing, clustering, and visualization tasks. These tools enabled a comprehensive and systematic analysis, ensuring the project's objectives were met with precision and efficiency.

A detailed list of the tools and libraries employed, along with their respective purposes, is provided below. Formal references to these resources can be found in the 'References' section.

1. **pandas**: Used for data manipulation and analysis, particularly for handling tabular data.
2. **numpy**: Utilized for numerical operations and array manipulation.
3. **matplotlib**: A visualization library for creating static, animated, and interactive plots.
4. **seaborn**: Built on matplotlib, it was used for advanced visualizations and aesthetic improvements.
5. **collections (Counter)**: Used to count occurrences of elements in datasets.
6. **scikit-learn**:
    - **StandardScaler**: For standardizing features by scaling them to a mean of 0 and a standard deviation of 1.
    - **PCA (Principal Component Analysis)**: Used for linear dimensionality reduction.
    - **KMeans**: Applied for clustering data into groups.
    - **GaussianMixture**: A clustering algorithm based on the Gaussian Mixture Model.
    - **TSNE (t-Distributed Stochastic Neighbor Embedding)**: Employed for visualizing high-dimensional data.
    - Metrics like **silhouette_score** were used to evaluate the quality of clusters.
7. **umap-learn**: A library for non-linear dimensionality reduction using UMAP (Uniform Manifold Approximation and Projection).

8. **mpl_toolkits.mplot3d (Axes3D)**: Facilitated 3D data visualization.

Each of these tools contributed significantly to different stages of the project, such as data preprocessing, dimensionality reduction, clustering, and visualization, making the analysis comprehensive and interpretable.

References

1. Bisong, E. (2019). Google Colaboratory. In *Building machine learning and deep learning models on Google Cloud Platform* (pp. 59–64). Springer. https://doi.org/10.1007/978-1-4842-4470-8
2. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
3. Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
4. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
5. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR, 31)*(3), 264–323. https://doi.org/10.1145/331499.331504
6. Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.
7. McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for dimension reduction [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1802.03426
8. Open Sourcing Mental Illness. (2016). *Mental health in tech survey 2016* [Data set]. Kaggle. https://www.kaggle.com/osmi/mental-health-in-tech-2016
9. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. Retrieved from https://scikit-learn.org
10. Python Software Foundation. (2023). *Python (version 3.10)* [Computer software]. Retrieved from https://www.python.org
11. Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581
12. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*(1), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
13. The pandas development team. (2023). *pandas: Python Data Analysis Library (version 1.5.0)* [Computer software]. Retrieved from https://pandas.pydata.org
14. Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021
15. Wilkinson, L., & Friendly, M. (2009). The history of the cluster heatmap. *The American Statistician, 63*(2), 179–184. https://doi.org/10.1198/tas.2009.0033
16. World Health Organization. (2022, June 17). Mental health: Strengthening our response. Retrieved from https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response