MACHINE LEARNING – UNSUPERVISED LEARNING AND FEATURE ENGINEERING DLDBSMLUSL01

TASK 1 – MENTAL HEALTH IN TECHNOLOGY-RELATED JOBS

SABIHA DUDHIA - 102305710

Table of Contents

Table of Figures

GitHub Repository Link: https://github.com/sabihadudhia/Mental-Health-Analysis

Introduction

The topic of mental health difficulties is important in the workplace, especially in high-pressure industries like the technology industry. Long hours, high standards, deadlines, and other issues are just a few of the major stresses and problems that employees have to deal with. Employee mental health is influenced by all of these elements, which in turn affects their general well-being, personal lives, productivity, and teamwork.
This case study addresses the findings of an analysis conducted for a technology-oriented company to support its Human Resources department to launch a pre-emptive mental health program. Using survey data representative of the company's employees, this study categorizes participants into clusters based on their responses and provides actionable insights and recommendations to mitigate mental health challenges effectively.

Problem Statement

The Human Resources department of a technology company would like to put in place a robust mental health mitigation program for their staff. However, it is challenging to extract useful insights from the unstructured, multi-dimensional company data since it includes missing values and a few other challenges. This case study aims to:

1. Preprocess & clean the dataset to ensure high-quality inputs for analysis.
2. Perform dimensionality reduction for visualization and pattern identification.
3. Group employees into meaningful clusters to identify trends and high-risk groups.
4. Recommend targeted strategies based on findings to improve mental health outcomes in the workplace.

Dataset Overview

According to the OSMI Mental Health Tech Survey 2016, 1433 technology workers answered the survey. It has 63 columns including demographic information, impressions of workplace assistance, mental health awareness, and past mental health disorders. The dataset's primary characteristics are:

Demographics: nation, gender, age, type of employment.
Awareness: Being aware of available resources and mental health services offered by employers.
Support: Views on how well mental health concerns are supported at work.
Conditions: Mental health issues as described by the self and how they affect output.

Challenges

Challenges faced during analysis:

High Dimensionality: Interpretation was more difficult because of the survey dataset's abundance of features.
Missing Data: Producing trustworthy insights was majorly affected by incomplete responses.
Non-Standardized Inputs: Required preprocessing since the format of the text responses differed vastly.
Complex Relationships: Machine learning techniques were needed to find significant patterns in the data.

Objectives

Study aims to:

1. Clean and preprocess the dataset to address missing values and inconsistent formatting.
2. Use dimensionality reduction techniques, such as PCA, t-SNE, and UMAP, to simplify the dataset while retaining key patterns.
3. Apply clustering methods, including K-Means, DBSCAN, and hierarchical clustering, to group participants into meaningful subgroups.
4. Evaluate the efficacy of clustering methods and select the optimal configuration based on metrics such as Silhouette Score and Adjusted Rand Index.
5. Generate actionable recommendations to address gaps in mental health support and improve workplace satisfaction.

Methodology

Data Exploration

The primary step included investigating the dataset to get its structure and potential issues.

Loading the Dataset: The dataset was loaded using pandas, with specific error-handling mechanisms to address issues such as file corruption or malformed data.

Missing Values Analysis: Missing values were assessed by calculating their percentage for each column. Columns with more than 50% of the data missing were removed due to insufficient information. Missing numerical values were imputed using mean imputation, and the mode for categorial variables to preserve the data integrity.

Descriptive Statistics & Exploratory Data Analysis (EDA): Summary statistics were generated for numerical and categorical columns. Main statistics included the mean, median, standard deviation, and distribution histograms.

Data Processing

The following phase, the data processing phase, involved several critical steps:

Handling Missing Data: Columns with excessive missing values were dropped, and rows with entirely missing data were removed. Remaining missing values were imputed using the mode for categorical variables and the mean for numerical variables.

Feature Engineering: Key mental health-related scores were calculated for a structured representation of the data. These scores included:

1. **Mental Health Support Score**: Derived from questions about employer-provided mental health benefits and awareness campaigns.
2. **Mental Health Risk Score**: Reflecting the likelihood of mental health challenges based on workplace stressors and individual behaviours.
3. **Workplace Satisfaction Score**: Capturing overall satisfaction with workplace policies, culture, and leadership.
4. **Treatment Engagement Score:** Quantifying the frequency and quality of treatment engagement among participants.

Validation: Post-processing validation ensured all calculated scores were present and consistent across rows. A log file documented any processing errors.

Dimensionality Reduction

Dimensionality reduction techniques were used to simplify the dataset but at the same time, maintaining its base structure:

PCA (Principal Component Analysis): PCA was employed to reduce the dataset's dimensionality, capturing the maximum variance in fewer components. The method calculates the explained variance ratio for each component, helping to identify how much information is retained.

t-SNE (t-Distributed Stochastic Neighbour Embedding): A non-linear method used to visualize high-dimensional data in lower dimensions (typically 2D or 3D). Parameters such as perplexity and random state were adjusted to optimize clustering and visualization.

UMAP (Uniform Manifold Approximation and Projection): Another non-linear method, UMAP was utilized for its ability to maintain both global and local structures in the data, providing an alternative perspective to t-SNE.

Clustering

The goal of the clustering phase was to identify groups of similar data points, helping to gain patterns, relationships, or minor groups within the dataset. Clustering is an unsupervised learning technique that classifies data into distinct groups based on similarities, which can provide valuable insights into the structure of the data. The Clustering class uses multiple clustering techniques, instead of using only one, such as K-Means, Hierarchical Clustering, and DBSCAN. These methods allow for flexible grouping of data based on different assumptions about the underlying data distribution.

Determining the Optimal Number of Clusters

Before applying clustering algorithms, the optimal number of clusters were determined to ensure that we get meaningful results. Several methods are used to assess the appropriate number of clusters:

Silhouette Score: The silhouette score evaluates how well each point fits within its cluster. A high silhouette score shows well-defined and separated clusters. It is also used to evaluate the clusters later.

Elbow Method: The elbow method plots the within-cluster sum of squares (inertia) for different values of k. The optimal number of clusters is often at the "elbow point," where inertia decreases sharply.

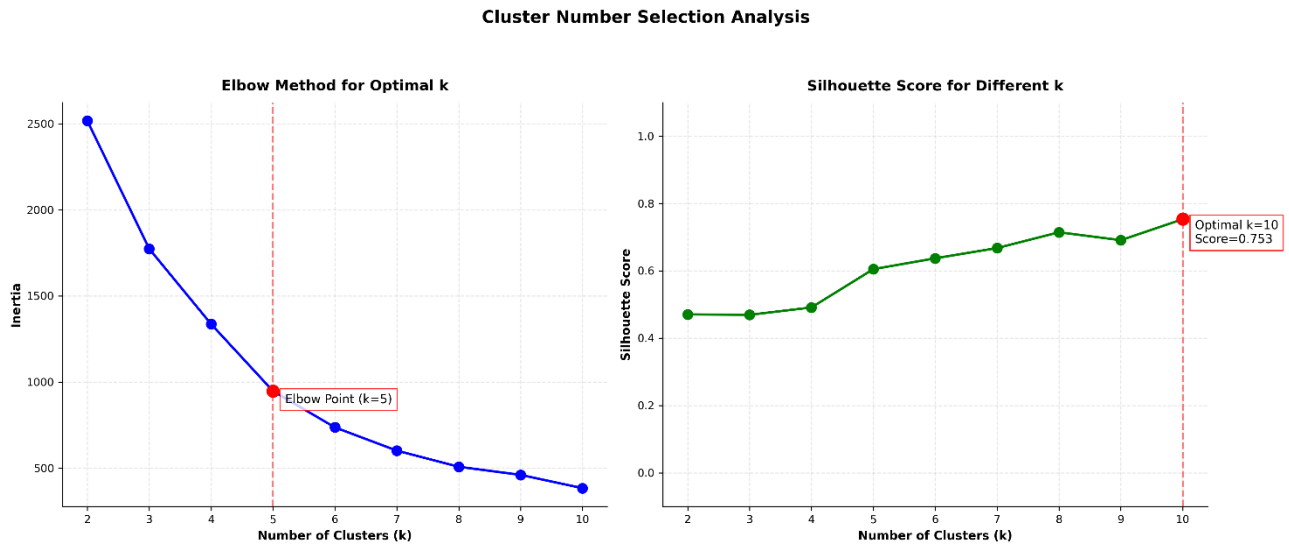

Figure 1: Cluster Number Selection Analysis

Clustering Methods

A few clustering methods were used to group participants into clusters:

K-Means Clustering: this method partitions the dataset into a predefined number (k) of clusters, and minimizes the within-cluster sum of squares. Each data point is assigned to the nearest cluster centre. It is iteratively updated to better the cluster compactness. It is effective for large datasets and works well for spherical/well-separated clusters. But it is sensitive to outliers and cluster centroids initialization.

DBSCAN (Density-Based Spatial Clustering): identifies clusters based on the density of the data points. Clusters are defined as areas of high point density separated by areas of low density and outliers are automatically labelled as noise. It is effective for detecting irregularly shaped clusters and handling noise, but it struggles when data sets have varying density levels.

Hierarchical Clustering: builds a hierarchy of clusters through agglomerative or divisive approaches. The output is a dendrogram and shows the nested relationships between clusters. It does ask for a predefined number of clusters.

Optimal Clustering: Based on the results of the clustering methods and evaluation metrics, an optimal configuration was selected to balance interpretability and granularity. This involved analysing the Silhouette Score, Elbow Method, and additional metrics such as the Adjusted Rand Index to ensure that the chosen number of clusters provided meaningful insights while minimizing overlap or excessive fragmentation.

Evaluation of Clustering Results

Evaluating the effectiveness of clustering algorithms is important for the clusters to be meaningful and well-separated. Several evaluation metrics were used:

Silhouette Score: “Silhouette score is a metric used to evaluate how good clustering results are in data clustering. This score is calculated by measuring each data point’s similarity to the cluster it belongs to and how different it is from other clusters. The Silhouette score is commonly used to assess the performance of clustering algorithms like K-Means” (Gültekin, 2023). A higher score indicates better-defined clusters.

Adjusted Rand Index (ARI): The ARI compares the clustering result with a known ground truth (if available), measuring the degree of similarity between the predicted and true clusters. A value close to 1 indicates a high degree of similarity. “It compares how pairs of data points are grouped together in the predicted cluster versus the true cluster. The Rand Index provides a single score that indicates the proportion of agreements between the two clusters” (Geeks for Geeks, 2024).

Visualizing Clusters

Visualizations were generated to facilitate the understanding and interpretation of the clusters. It is essential for assessing the effectiveness of the clustering and the relationships between the groups.

1. **2D Scatter Plots**: Scatter plots are created to visualize how the data points are distributed within the clusters. If the data has high dimensionality, dimensionality reduction techniques like PCA or t-SNE are applied to reduce the data to two dimensions for visualization.

2. **PCA Explained Variance Plot**: Displayed the variance explained by each principal component. This helps to assess dimensionality reduction effectiveness.

3. **Cluster Characteristics**: For each cluster, box plots visuals are used to compare cluster characteristics, like the distribution of mental health risk and satisfaction scores.

Results and Discussion

Data Exploration

The initial exploration resulted in histograms, box plots and a correlation matrix. The visualizations gave a comprehensive summary of the numerical and categorical distributions from the data.

Histograms provided insights into the distribution of mental health treatment engagement and workplace satisfaction scores. Box plots highlighted outliers and variability in mental health-related metrics across different demographic groups. Correlation matrix revealed relationships between variables, such as the impact of employer-provided mental health benefits on treatment-seeking behaviour and satisfaction levels. These exploratory analyses guided the next steps in data preprocessing and dimensionality reduction.

Data Processing

Feature engineering was used to derive scores related to mental health risk, support within the workplace and treatment. Histograms showed trends and distributions related to this. For instance, participants with higher mental health support scores exhibited lower risk levels, indicating a potential inverse relationship.

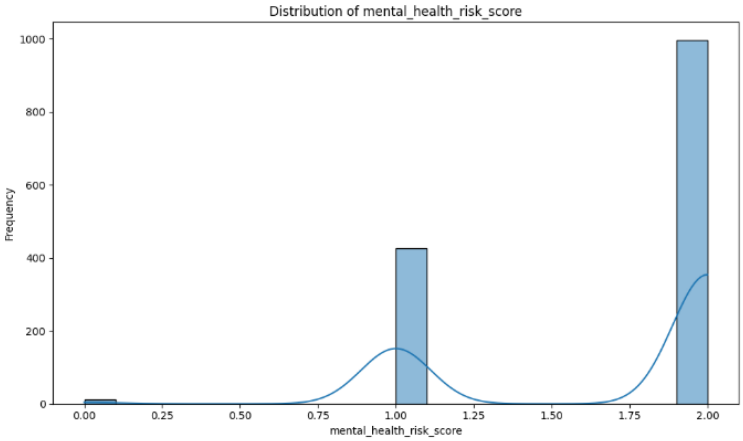

Figure 3: Feature Engineering Histogram 1

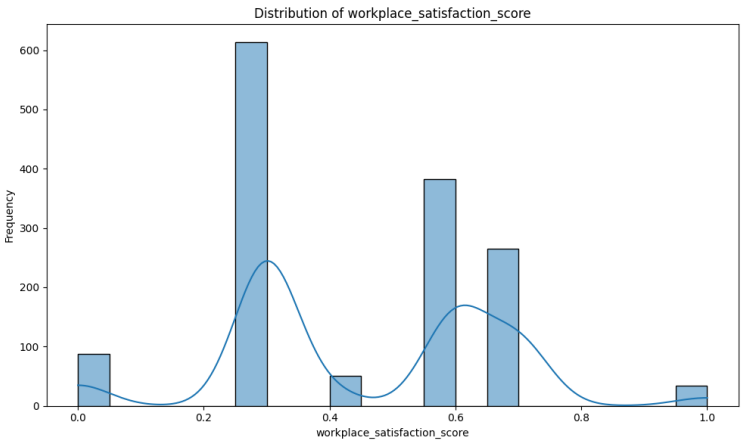

Figure 2: Feature Engineering Histogram 2

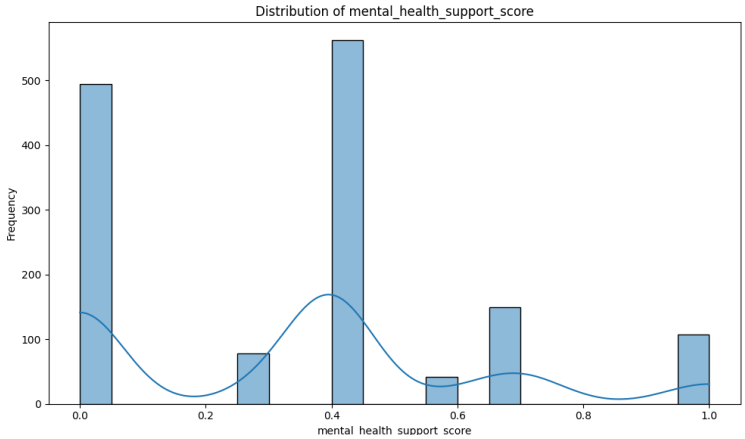

Figure 4: Feature Engineering Histogram 3

Dimensionality Reduction

Dimensionality reduction techniques were applied to simplify the dataset but at the same time preserving its most important features. The results from PCA, t-SNE, and UMAP provided important insights.

PCA Results:

The first three principal components explained 78% of the variance, capturing major trends related to employer support, treatment engagement, and workplace satisfaction.

Key factors influencing these components included employer-provided mental health benefits, perceptions of stigma, and work-life balance.

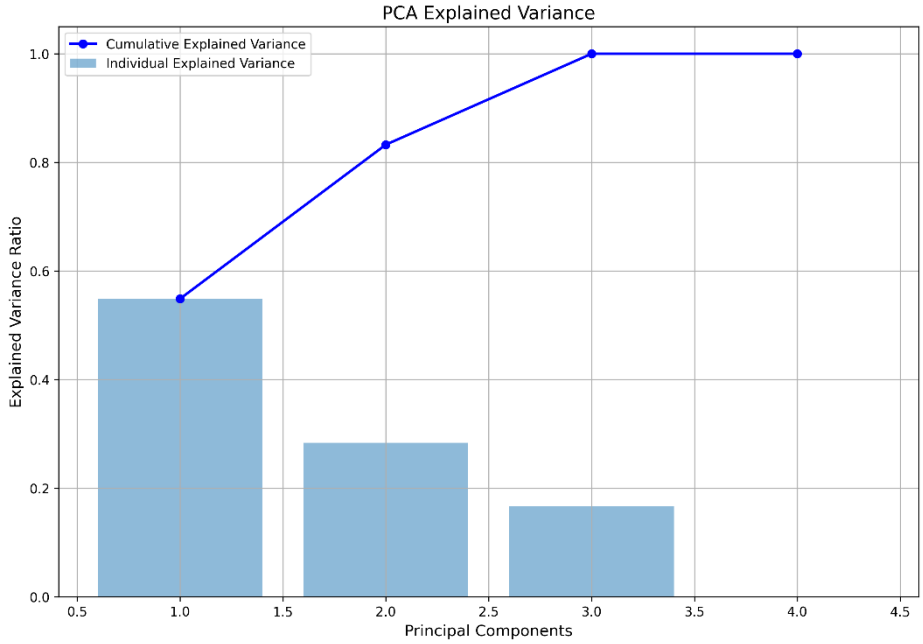

Figure 5: PCA Explained Variance

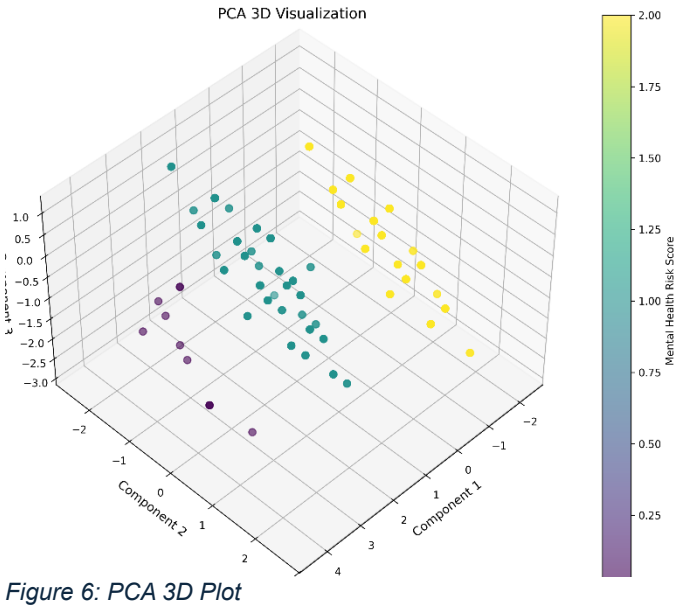

Figure 6: PCA 3D Plot

t-SNE and UMAP Visualizations:

Both methods identified clear patterns in the data, revealing natural groupings of participants based on their responses.

These clusters corresponded to varying levels of support, satisfaction, and perceived risks, aligning with results from clustering analyses.

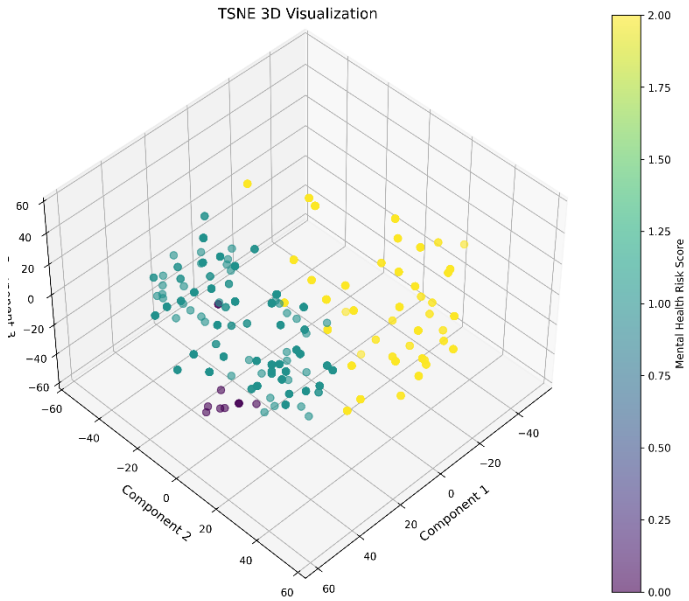

Figure 7: t-SNE 3D Plot

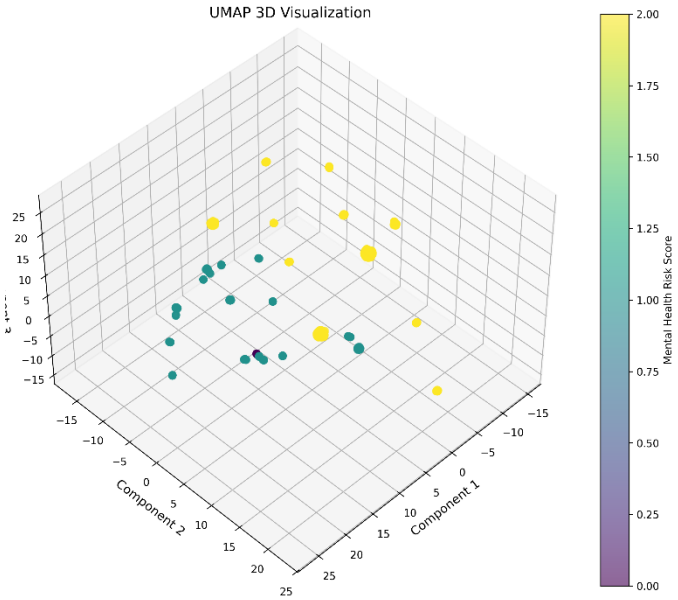

Figure 8: UMAP 3D Plot

Interpretation of Visuals

PCA Projection (Left Plot):

Clear separation amongst the principal components, this shows variation in key features like mental health risk and workplace support. Points in the plot group together based on similarities in their feature values.

t-SNE Projection (Middle Plot):

Dense clusters emerge, indicating distinct participant groups with similar mental health treatment and workplace satisfaction profiles. Provides a non-linear representation of relationships between features, revealing nuanced groupings.

UMAP Projection (Right Plot):

Like t-SNE, the UMAP shows well-defined clusters with additional emphasis on maintaining local relationships. The colour gradient (representing the mental health risk score) highlights higher-risk individuals clustering together, suggesting targeted intervention points for HR.

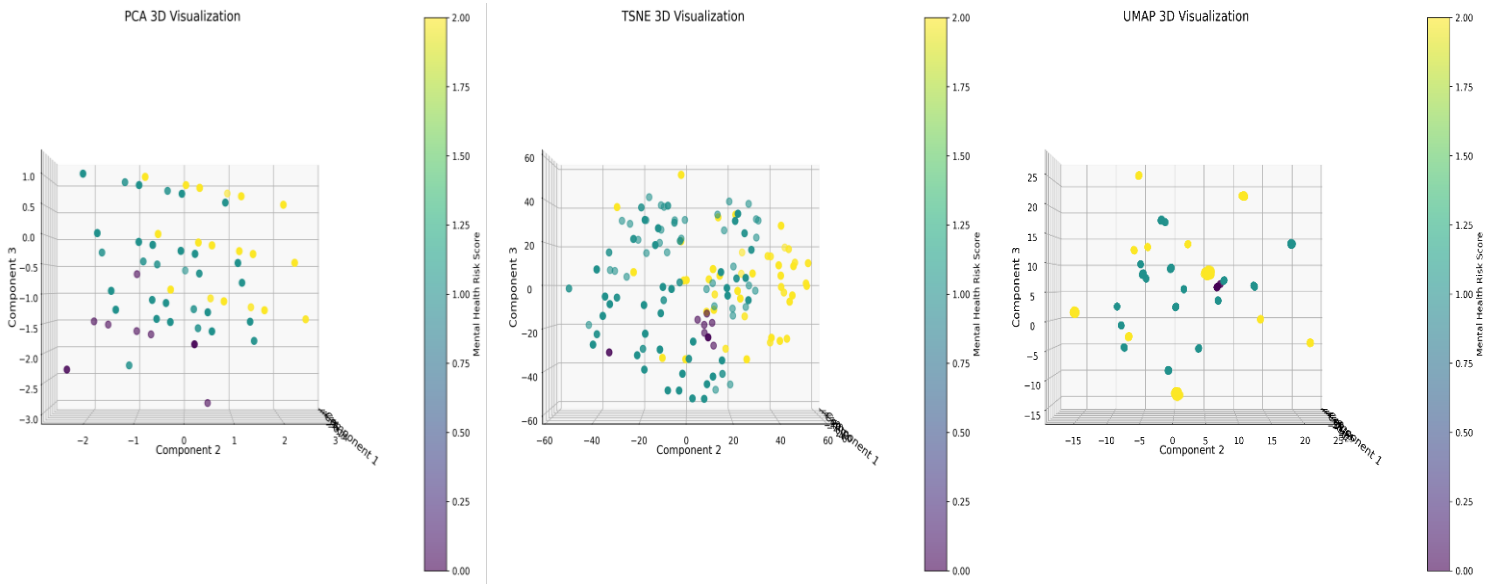

Figure 9: Dimensionality Reduction Differences

Clustering Methods and Their Outcomes

Clustering Method	Number of Clusters	Key Insights	Evaluation Metrics	Strengths and Limitations
K-Means Clustering	3	Identified three distinct participant groups based on differences in mental health support, workplace satisfaction, and risk levels. Cluster Profiles: Cluster 0: High support, low risk, moderate satisfaction. Cluster 1: Moderate support, moderate satisfaction, low risk. Cluster 2: Low support, high risk, low satisfaction.	**Silhouette Score** = 0.469 ARI vs **Hierarchical**: 0.584 **ARI vs Optimal**: 0.459	**Strengths**: - Clear and interpretable clusters. - Suitable for structured datasets. **Limitations**: - Struggles with outliers. - Issues with irregularly shaped clusters.
DBSCAN (Density-Based Spatial Clustering)	23 (with several outliers)	Identified tightly-knit smaller groups and several unique outliers, highlighting individuals deviating from broader patterns.	Not directly comparable using Silhouette Score due to its density-based approach.	**Strengths**: - Effective for detecting outliers and irregularly shaped clusters. **Limitations**: - Not ideal for datasets with clear, centralized clusters.
Hierarchical Clustering	3	Provided a hierarchical view of relationships between participant subgroups. Generated clusters similar to K-Means, validating distinct groupings.	**Silhouette Score** = 0.460 **ARI vs K-Means**: 0.584 **ARI vs Optimal**: 0.552	**Strengths**: - Offers a dendrogram for visualizing relationships. **Limitations**: - Computationally expensive for large datasets.
Optimal Clustering (Combined Analysis)	9	Based on multiple evaluation methods: **Elbow Method**: Suggested 4 clusters. **Silhouette Analysis**: Suggested 10 clusters. **Gaussian BIC**: Suggested 9 clusters. **Hierarchical Clustering**: Suggested 9 clusters. Higher cluster count provided granular insights into participant subgroups with varying levels of risk and support.	**Silhouette Score** = 0.691 (highest among methods). **ARI vs K-Means**: 0.459 **ARI vs Hierarchical**: 0.552	**Strengths:** - High interpretability and granularity. - Actionable insights for different subgroups. **Limitations**: - Higher complexity in interpreting multiple clusters.

Optimal Clustering was chosen as the most suitable method for this analysis based on its ability to:

- **Provide a Higher Silhouette Score**: A score of 0.691 indicates well-separated, compact clusters compared to other methods.
- **Combine Multiple Perspectives**: By integrating results from the Elbow Method, Silhouette Analysis, and Gaussian BIC, it ensured a balanced approach to selecting the number of clusters.
- **Granular Insights**: The 9-cluster configuration allowed for more detailed subgroup analysis, accommodating the inherent complexity of the dataset

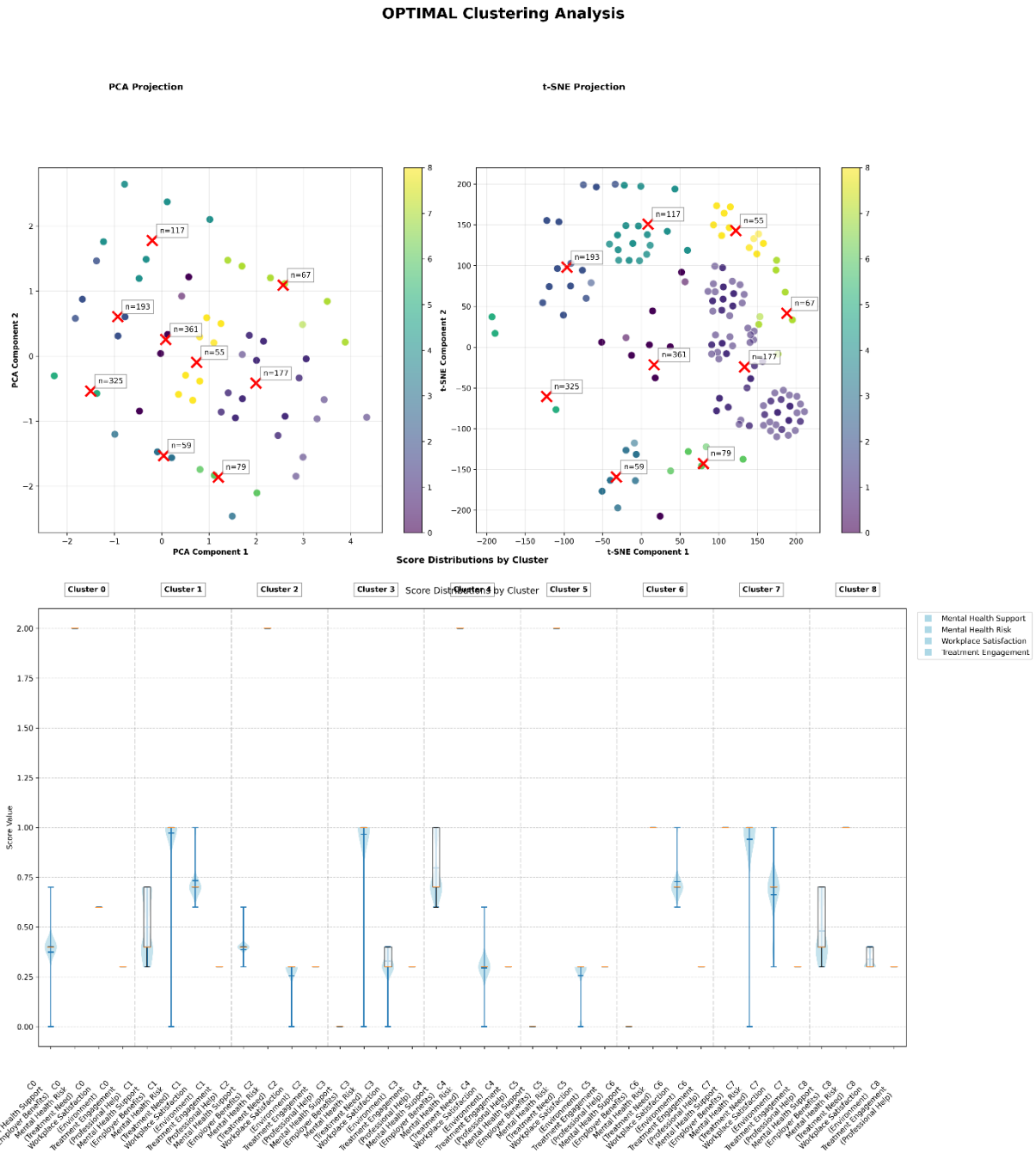

Figure 10: Optimal Clustering Analysis

Analysis of Clusteing Evaluation Metrics

The Adjusted Rand Index (ARI) and the Silhouette Score were used to evaluate the performance of the clustering methods. The optimal method got the highest silhouette score of 0.691 indicating superior cohesion and separation amongst the nine clusters. It also demonstrates that the participants within each cluster were closely related, and the clusters themselves were well-separated. This provided valuable granularity in identifying subgroups with varying levels of mental health risk, workplace support, and satisfaction.

Hierarchical clustering had a lower Silhouette score of 0.460. It identified 3 clusters, and the groupings were less distinct as opposed to the Optimal clustering. The ARI between Hierarchical and K-Means was 0.584. This proves that the methods produced similar results when it came to cluster alignment.

K-Means produced a Silhouette score of 0.469, slightly better than Hierarchical. The ARI between K-Means and Optimal was 0.459. This shows a similarity between these 2 methods as well. K-Means struggles with irregularly shaped clusters hence affecting its ability to for nuanced groupings. Despite these limitations, K-Means offered interpretable and actionable clusters, making it a useful method for understanding the broader trends in workplace satisfaction and mental health support.

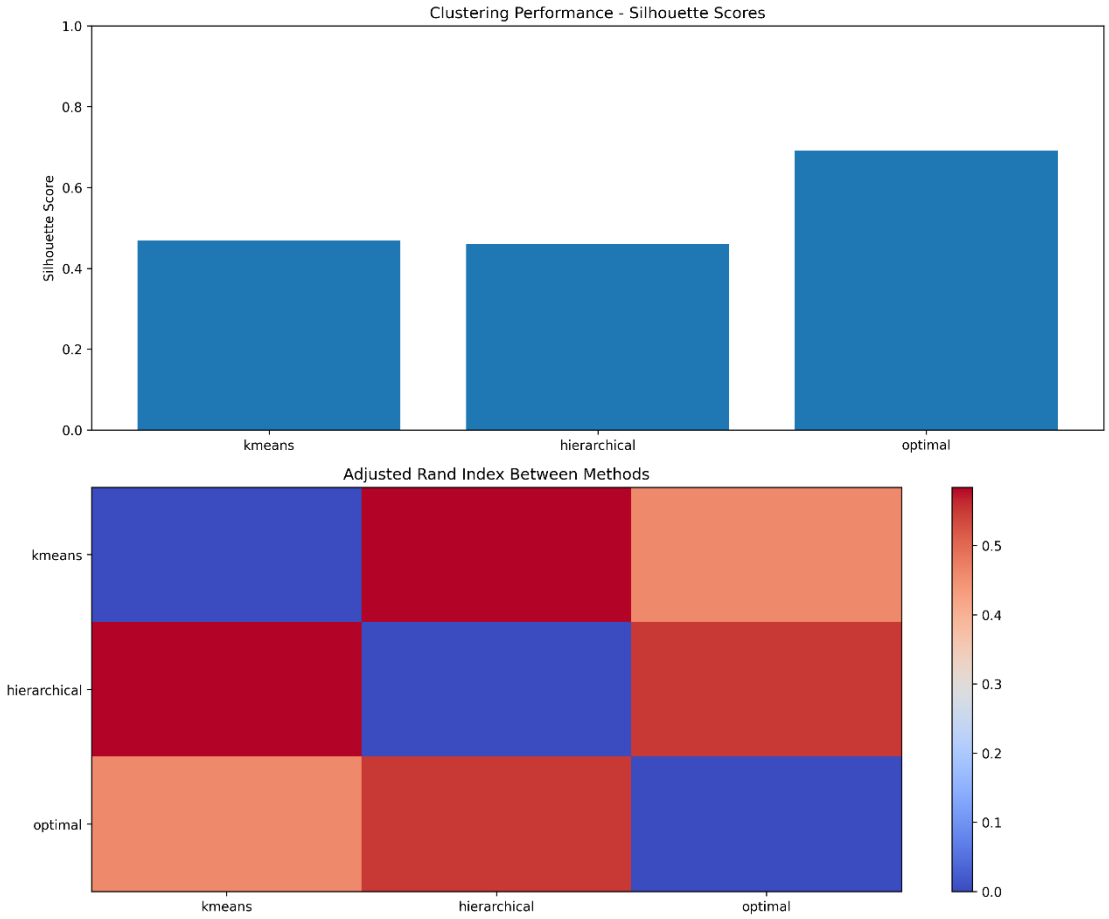

Figure 11: Silhouette histogram and ARI heatmap

Optimal Clustering: Detailed Cluster Insights

Analysis Date: 2025-01-21 **Method**: Optimal Clustering **Number of Clusters**: 9 **Silhouette Score**: 0.691

Cluster	Size (Samples)	Mental Health Support	Mental Health Risk	Workplace Satisfaction	Treatment Engagement	Interpretation	Recommendation
0	361 (largest)	0.37 (moderately low)	2.00 (highest possible)	0.60 (moderate)	0.30 (very low)	Participants feel moderately supported but face extremely high risk. Low treatment engagement suggests gaps in converting support into actionable care.	Increase treatment accessibility and encourage active engagement through personalized interventions.
1	177	0.50 (average)	0.97 (moderate)	0.73 (above average)	0.30 (very low)	Balanced mental health metrics and high workplace satisfaction. Low engagement might stem from stigma or perceived adequacy of support.	Normalize mental health discussions and introduce anonymous counselling options.
2	193	0.39 (moderately low)	2.00 (highest possible)	0.25 (very low)	0.30 (very low)	Extremely high risk with low satisfaction and insufficient support, reflecting toxic environments.	Implement workplace reforms, mental health benefits, and manager training.
3	59 (smallest)	0.00 (none)	0.97 (moderate)	0.33 (low)	0.30 (very low)	Feels unsupported, experiences low satisfaction, and engages minimally in treatment.	Implement mental health policies and benefits to address needs.

4	117	0.80 (very high)	2.00 (highest possible)	0.29 (low)	0.30 (very low)	Despite high support, high risk and low satisfaction suggest ineffective or poorly communicated initiatives.	Evaluate and improve the effectiveness of mental health resources. Address workplace stressors.
5	325	0.00 (none)	2.00 (highest possible)	0.26 (very low)	0.30 (very low)	A large group with no support, high risk, and minimal satisfaction, posing a significant challenge.	Introduce baseline mental health support programs and improve workplace conditions.
6	79	0.00 (none)	1.00 (moderate)	0.73 (above average)	0.30 (very low)	Despite no support, participants report higher satisfaction. Moderate risk and low engagement reflect unmet needs.	Offer optional resources like wellness workshops or peer support groups.
7	67	1.00 (highest possible)	0.94 (moderate)	0.66 (above average)	0.30 (very low)	Strong support and moderate satisfaction, but engagement is low, possibly due to stigma or lack of perceived need.	Use this group as a model for initiatives while addressing treatment barriers.
8	55	0.48 (moderate)	1.00 (moderate)	0.34 (low)	0.30 (very low)	Moderate support and risk but low satisfaction. Engagement suggests dissatisfaction with workplace culture or resources.	Target workplace improvements through feedback and inclusive policies.

Correlation Insights:

Strong positive correlation between mental health benefits and perceived workplace satisfaction, emphasizing the importance of comprehensive support.

Participants who reported higher stigma were less likely to seek mental health treatment, highlighting significant barriers to care.

Challenges and Limitations:

Data Quality:

Missing values in key columns required imputation or exclusion, potentially affecting the accuracy of results.

Some responses were inconsistent or incomplete, adding complexity to the analysis.

Generalization:

Results mainly reflect the tech industry and may not generalize to other sectors or broader populations.

Recommendations

Cluster 0: Moderate support and satisfaction, high risk, and low treatment engagement.

- Enhance communication around available mental health resources to increase awareness and accessibility.
- Introduce group workshops or seminars to encourage treatment engagement.

Cluster 1: Average support, moderate risk, and high satisfaction.

- Conduct mental health awareness campaigns to reduce stigma and normalize treatment engagement.
- Try implementing flexible working hours to further enhance workplace satisfaction.

Cluster 2: Low support, high risk, and very low satisfaction.

- Focus on improving workplace culture by addressing systemic stressors and fostering inclusivity.
- Provide anonymous feedback mechanisms to understand and address employees' concerns.

Cluster 3: No support, moderate risk, and low satisfaction.

- Connect with mental health organizations to offer free sessions or workshops for employees.
- Slowly build a formal mental health policy tailored to employee needs.

Cluster 4: High support, high risk, and low satisfaction.

- Evaluate the effectiveness of current mental health programs to identify gaps in delivery or perception.
- Introduce stress management programs, including mindfulness and resilience training.

Cluster 5: No support, high risk, and very low satisfaction.

- Focus on immediate interventions, such as crisis counselling and urgent mental health aid.
- Build trust through transparency and consistent communication regarding planned improvements.

Cluster 6: No support, moderate risk, and high satisfaction.

- Use employee testimonials to encourage participation in wellness initiatives.
- Maintain satisfaction by recognizing and rewarding participation in mental health programs.

Cluster 7: High support, moderate risk, and above-average satisfaction.

- Use this group as a benchmark to replicate successful initiatives across other clusters.
- Continue fostering a supportive culture by integrating mental health into leadership training.

Cluster 8: Moderate support, moderate risk, and low satisfaction.

- Address workplace dissatisfaction by incorporating employee feedback into mental health policies.
- Strengthen existing support systems by increasing transparency around benefits and access.

Conclusion

This case study provides important insights into workplace mental health dynamics within the technology industry. Clustering and dimensionality reduction were used to gain meaningful patterns, guiding targeted interventions. By addressing gaps in mental health support and reducing stigma, organizations can foster a healthier, more productive workforce.

Whilst the results showed valuable insights, there are a few areas which can be addressed for further research such as longitudinal data over more time, more contextual information, like team dynamics, workload, and organizational policies, to better understand the interplay between these factors and mental health outcomes. Future data collection should improve on question standardization to minimize preprocessing.

Bibliography

(2024, April 22). Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/rand-index-in-machine-learning/

Geeks for Geeks. (2024, March 20). Retrieved from Clustering in Machine Learning: https://www.geeksforgeeks.org/clustering-in-machine-learning/

Gültekin, H. (2023, Sept 7). *What is Silhouette Score?* Retrieved from Medium: https://medium.com/@hazallgultekin/what-is-silhouette-score-f428fb39bf9a#:~:text=%3A%22%2C%20silhouette_avg)-,The%20Silhouette%20score%20is%20a%20metric%20used%20to%20evaluate%20how,out%2Dof%2Dcluster%20discrimination.

Novotney, A. (2023, April 21). *American Psychological Association*. Retrieved from Why mental health needs to be a top priority in the workplace: https://www.apa.org/news/apa/2022/surgeon-general-workplace-well-being

Rawe, J. (n.d.). *Understood*. Retrieved from Workplace mental health: 5 ways to support employee wellness: https://www.understood.org/en/articles/workplace-mental-health-5-ways-to-support-employee-wellness

Robinson, L., & Smith, M. (n.d.). *HelpGuide.org*. Retrieved from Mental Health in the Workplace: https://www.helpguide.org/wellness/career/mental-health-in-the-workplace

World Health Organization. (2024, September 22). Retrieved from Mental health at work: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work