**Mental Health Clustering based on Tech Jobs**

Case Study for Course DLBDSMLUSL01

Machine Learning – Unsupervised Learning and Feature Engineering

20.11.2024

Mohsen Nasiri

Matriculation Number: 32206338

Tutor’s Name: Christian Müller-Kett

**Table of Contents**

1. Introduction
2. Methodology  
   2.1 Data Preprocessing  
   2.2 Dimensionality Reduction  
   2.3 Clustering
3. Results and Analysis  
   3.1 Cluster Characteristics  
   3.2 Evaluation Metrics
4. Conclusion
5. References
6. Appendices

**1. Introduction**

This study analyzes survey data from technological jobs to find clusters of employees based on their mental health responses and derive actionable insights for Human Resources (HR) to develop targeted mental health programs.

**Context:**

The HR department of a tech company is initiating a program to address mental health issues among its employees. To support this effort, a survey conducted among technology professionals was analyzed, representing the organizational workforce. The survey aimed to capture several factors influencing mental health, workplace dynamics, and individual experiences.

The dataset presents significant challenges, including high dimensionality, complexity, and the presence of missing values and non-standardized textual inputs. These factors make direct interpretation and application of insights difficult.

**Objectives:**

The objective is to simplify the dataset while retaining its core characteristics, providing actionable insights to guide HR initiatives by categorizing survey participants and creating meaningful visualizations and clusters.

**Key points:**

1. Reduce the complexity of the data while retaining key insights for actionable understanding.
2. Categorize survey participants based on their responses related to mental health and workplace factors.
3. Provide cluster-specific insights to aid HR in designing targeted interventions and improvements.

These clusters and their key characteristics will help identify potential leverage points for targeted measures, ultimately supporting a healthier workplace environment.

**1. Libraries**

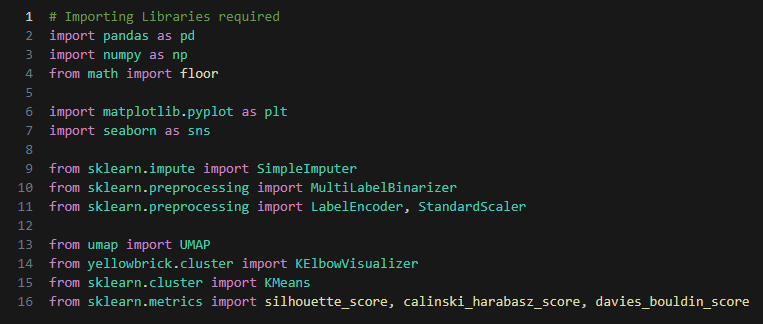
The Pandas library was utilized for reading, cleaning, and manipulating the dataset and loading the csv file into a data frame. And for further handling arrays and other operations related to the data frame, Numpy was used.

For visualizing and plotting data and trends, Matplotlib was implemented along Seaborn which is built on Matplotlib to enhance the visualizations.

During the data preprocessing, Sklearn Impute Manages missing data by filling in missing values using strategies like mean, median, or most frequent value. Also, during the encoding and scaling the data, Sklearn Preprocessing provided MultiLabelBinarizer which encodes categorical variables with multiple labels into binary indicators, LabelEncoder that encodes categorical labels into numeric values, and StandardScaler to standardize features by removing the mean and scaling them to unit variance.

For dimensionality reduction UMAP was used. However, in the early versions of the code PCA from Sklearn Decomposition was implemented which was not optimal.

KelbowVisualizer from Yellowbrick was utilized to determine the optimal number of clusters for using the elbow method.

For clustering, the KMeans algorithm partitions the dataset based on similarity. To evaluate the clustering, metrics such as Silhouette Score, Calinski Harabasz Score, and Davies Bouldin Score are used. These metrics were implemented using Scikit-learn referenced as Sklearn.

**2. Simplifying**

The dataset initially contained inconsistently named columns, varied text formatting, and ambiguous categorical values. To improve usability and simplicity the following steps were implemented.

The original column names were complex and unclear. Each column was renamed to a more concise and descriptive format that aligns with the data it represents.

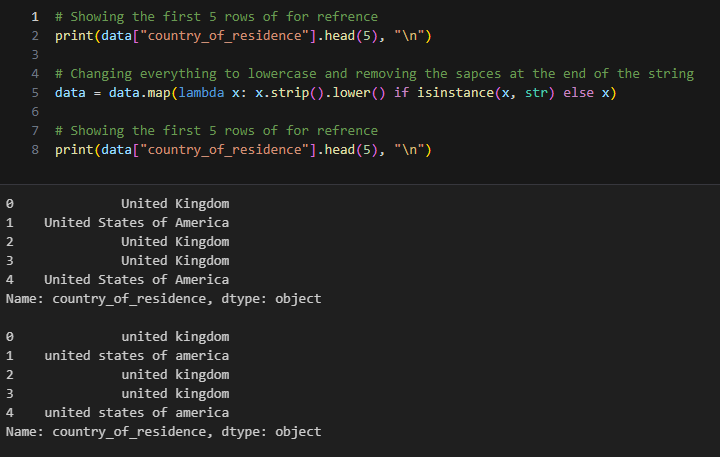
**A screen shot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generated**

All text data was converted to lowercase to ensure consistency in processing. Extra space at the ends of strings was also removed to avoid errors.

**A screen shot of a computer program

Description automatically generated**

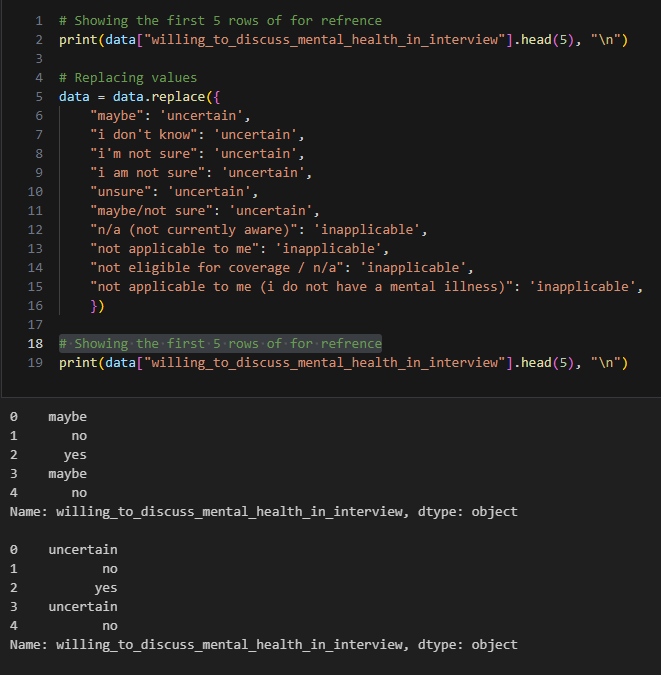
**3. Remapping Values in the Dataset**

Remapping was essential to standardize and simplify the categorical data in the dataset. It ensured consistency across responses, particularly where multiple terms expressed the same idea or where values needed to be separated and transformed for machine learning compatibility.

3.1. General Value Replacement in the Entire Dataset

In all columns with textual and categorical data, ambiguous or uncertain responses such as "maybe," "I don't know," and "unsure" were replaced with "uncertain". Additionally, non-applicable entries like "not applicable to me" were replaced with "inapplicable".  
This standardization reduced redundancy and ensured consistency, facilitating reliable categorical analysis.

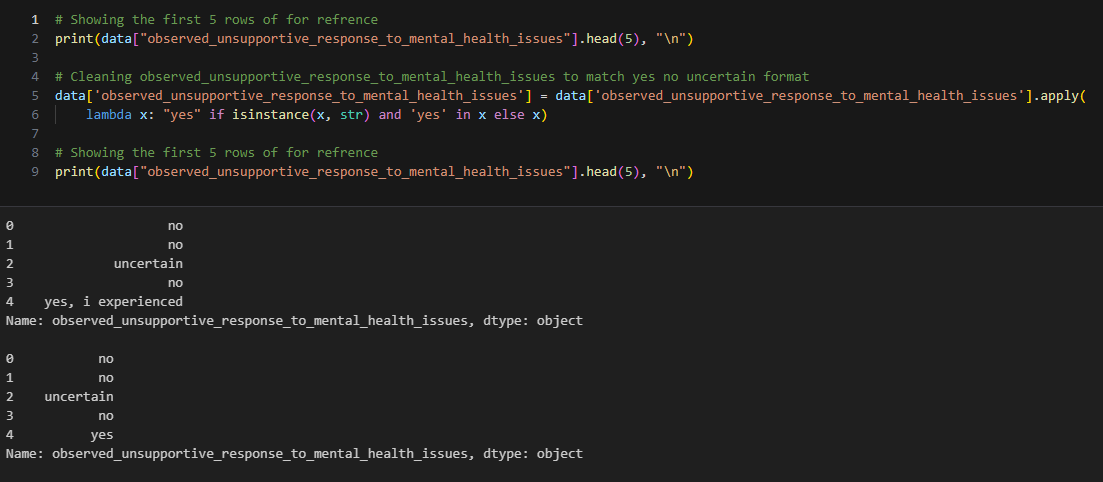
A screenshot of a computer program

Description automatically generated

3.2. Cleaning the observed\_unsupportive\_response\_to\_mental\_health\_issues Column

Responses containing variations of "yes" (e.g., "yes, definitely") were simplified to "yes" to maintain binary consistency.

A screen shot of a computer program

Description automatically generated

3.3. Standardizing Ranges in percentage\_of\_work\_time\_impacted\_by\_mental\_health

Text ranges like "1-25%" were mapped to their numerical equivalents (e.g., "1-25%" → 0.25).  
This conversion ensured compatibility with statistical and clustering techniques.

3.4. Unifying Formats in fear\_negative\_view\_from\_coworkers\_due\_to\_mental\_health and fear\_career\_impact\_from\_mental\_health\_disclosure

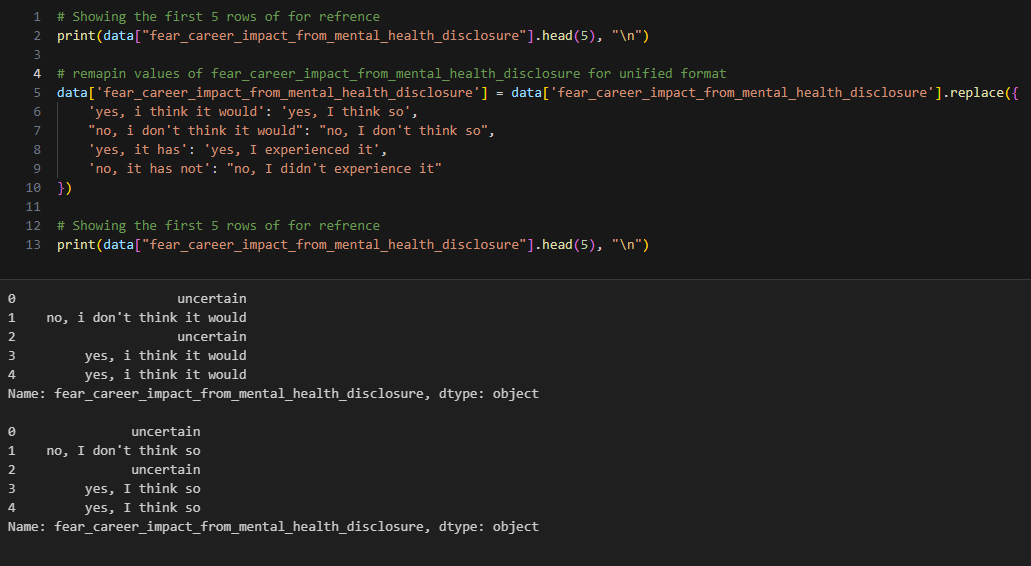
Responses such as "yes, I think they would" and "yes, they do" were remapped to "yes, I think so" and "yes, I experienced it," respectively.  
Negative responses like "no, I don’t think they would" were standardized to "no, I don’t think so."  
These adjustments ensured consistent phrasing to accurately reflect the intensity and type of response while maintaining uniformity in interpreting perceived or actual impacts.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

3.5. Simplifying Multi-level Responses Across Columns

For all categorical columns with multilevel responses, a mapping was applied to group similar terms into unified categories:

* "all," "yes," "always" → "all"
* "some," "sometimes" → "some"
* "none," "no," "never" → "none"

This simplification enabled ordinal encoding while preserving the hierarchical structure of the responses.

3.6. Cleaning the gender Column

Variations of "male" (e.g., "m," "mail") were replaced with "man." Similarly, variations of "female" (e.g., "f," "woman") were replaced with "woman."  
Entries that did not fit into these categories were labeled as "diverse."  
This standardization ensured inclusivity while maintaining consistency across gender data.

3.7. Addressing Missing States in us\_state\_or\_territory\_of\_residence and us\_state\_or\_territory\_of\_work

Missing or null values in location columns were replaced with "not specified" to account for individuals outside the United States or missing data.  
This approach ensured no null values remained, minimizing errors during analysis.

3.8. Standardizing Employee Ranges in number\_of\_employees

Text ranges like "more than 1000" were remapped to "above 1000," and ranges such as "500-1000" were rewritten as "500 to 1000."  
These adjustments addressed formatting issues (e.g., Excel interpreting these ranges as dates) while maintaining data integrity for analysis.

3.9. Cleaning reason\_for\_willingness\_to\_discuss\_physical\_health and reason\_for\_willingness\_to\_discuss\_mental\_health

Responses were grouped into broader categories for simplicity and consistency:

* Terms related to stigma or acceptance → "stigma"
* Privacy-related terms → "wont\_discuss"
* Open and transparent discussions → "honesty"
* Other responses → "other"  
  These categorizations reduced variability and clarified the analysis of employees' willingness to discuss health issues.

3.10. Cleaning diagnosed\_conditions, suspected\_conditions, and diagnosed\_conditions\_by\_professional

These columns contained varied textual descriptions of mental health conditions. To simplify analysis, conditions were grouped into broader categories:

Categories Defined

1. Anxiety and Mood Disorders: Includes generalized anxiety disorder, depression, and seasonal affective disorder.
2. Neurodevelopmental and Behavioral Disorders: Includes ADHD, autism spectrum disorders, and related impairments.
3. Trauma and Stress-Related Disorders: Includes PTSD, burnout, and dissociative disorders.
4. Substance and Addiction Disorders: Includes substance use and sexual addiction disorders.
5. Personality and Psychotic Disorders: Includes borderline personality disorder and schizophrenia.
6. Sexual and Gender Identity Disorders: Includes gender dysphoria and related conditions.
7. Neurological Disorders: Includes traumatic brain injuries and tinnitus.
8. Crisis and Mental Health Conditions: Includes suicidal ideation.
9. Other: Includes sleep and eating disorders.

Example Transformations

* Original Entry: "depression | anxiety disorder (generalized, social, phobia, etc)"
* Transformed Entry: ["anxiety\_and\_mood\_disorders", "anxiety\_and\_mood\_disorders"]

By grouping similar conditions into meaningful categories, the dataset was simplified without losing critical distinctions.  
This enhanced compatibility with clustering and visualization techniques, allowing for actionable insights tied to broader mental health domains rather than specific conditions.

3.11. Separating Work Positions

The work\_position column was standardized by splitting entries containing multiple roles into lists of individual roles. This transformation ensured each role was treated as a separate entity for analysis.

Impact of Remapping

These transformations streamlined the dataset, improving its quality and making it more interpretable. The remapping ensured consistent patterns across similar responses, which is critical for clustering algorithms and other machine learning techniques.

**4. Preprocessing the Dataset**

Preprocessing was a crucial step to prepare the dataset for effective analysis and machine learning application. This phase focused on handling missing values, encoding categorical data, and scaling numerical features. These steps ensured that the dataset was structured, clean, and ready for dimensionality reduction and clustering.

**4.**1. Handling Missing Values

Missing data can significantly affect the performance of machine learning algorithms. The following strategies were employed:

* Numerical Columns: Missing values were replaced with the median of the respective column. This approach was chosen to minimize the influence of outliers.
* Categorical Columns: Missing values in categorical features were imputed with the most frequent category. This ensured that the data distribution remained consistent.

**4.**2. Removing Outliers

* Age Column: The dataset included outliers in the "age" feature, such as values outside the realistic working age range. To address this, entries below 18 and above 80 were removed or marked as missing.
* The cleaned "age" feature was then grouped into bins (e.g., 18–30, 30–40) for better interpretability and analysis.

**4.**3. Encoding Categorical Variables

The KMeans algorithm works with numerical data. Therefor the categorical variables were converted using the following methods:

* Binary Encoding: Features with two possible values (e.g., "yes" or "no") were encoded as 1 and 0.
* Ordinal Encoding: For features with an inherent order (e.g., "none," "some," "all"), numerical values were assigned based on the hierarchy.
* Multi-label Binarization: Columns with multiple categories (e.g., "work positions") were expanded into separate binary columns, each representing a single category.

**4.**4. Simplifying Complex Features

Some columns contained detailed information, such as mental health conditions or job roles, which were challenging to analyze directly. These were simplified using the following methods:

* Mental Health Conditions: Diagnoses and conditions were categorized into broader groups (e.g., "anxiety and mood disorders," "neurodevelopmental disorders") to facilitate meaningful analysis.
* Work Positions: Job roles were split into individual categories and encoded as binary indicators.

**4.**5. Scaling Numerical Features

Numerical columns were standardized to ensure that features with larger values did not dominate the clustering algorithms. This was achieved by scaling all numerical values to have a mean of 0 and a standard deviation of 1.

Impact of Preprocessing

These preprocessing steps addressed the challenges of missing data, high cardinality, and feature variability in the dataset. By ensuring that the data was clean and well-structured, preprocessing laid the groundwork for accurate clustering and visualization while preserving the integrity of the original information.

5. Encoding the Dataset

Encoding was necessary to transform categorical data into numerical formats suitable for machine learning algorithms. Different strategies were applied based on the type of categorical data and its complexity. Below is an explanation of how encoding was performed.

5.1. Binary Encoding

For columns with two distinct categories (e.g., "yes" and "no"), binary encoding was applied.

* Responses were mapped to numerical values: "yes" as 1 and "no" as 0.
* This method simplified binary columns, making them immediately usable for clustering.

5.2. Ordinal Encoding

For columns with ordered categories (e.g., "none," "some," "all"), ordinal encoding was used to assign values based on the natural order of the categories. For example:

* "none" = 0, "some" = 1, "all" = 2.
* In cases where categories like "uncertain" or "inapplicable" appeared, these were assigned distinct values (e.g., -1) to differentiate them as special cases.

5.3. Multi-label Binarization

Columns with multiple categories, such as job roles or mental health conditions, required a different approach:

* Each unique category was transformed into a binary column, where 1 indicated the presence of a category for a particular row and 0 indicated its absence.
* This method allowed for multi-dimensional representation without losing information.

5.4. Mapping Specialized Columns

Some columns required custom mappings due to their complexity. For example:

* Disclosure Preferences: Categories such as "no, because it doesn’t matter," "sometimes, if it comes up," and "yes, always" were mapped to numerical values (0, 2, and 3, respectively) based on their logical order.
* Fear of Career Impact: Degrees of fear (e.g., "uncertain," "no, I don’t think so," "yes, I think so") were mapped to an ordinal scale (-1, 0, 2), reflecting their intensity.

5.5. Frequency Encoding

For features with high cardinality and no clear ordering (e.g., reasons for willingness to discuss mental health), frequency encoding was applied:

* Categories were replaced with their frequency counts within the dataset.
* This approach retained the significance of popular categories without expanding the dimensionality.

5.6. Multi-dimensional Representation

Some columns, such as mental health conditions, were grouped into broader categories (e.g., "anxiety and mood disorders," "trauma and stress-related disorders"). This reduced complexity and provided meaningful insights.

Impact of Encoding

Encoding transformed raw categorical data into a structured numerical format, making it suitable for clustering algorithms. The approach balanced the preservation of information with the reduction of complexity, ensuring that key patterns in the data were not lost during analysis.

6. Scaling, Dimensionality Reduction, and Clustering

Scaling, dimensionality reduction, and clustering were critical steps to simplify the dataset while retaining meaningful patterns. These methods reduced computational complexity, improved visualization, and enabled clear identification of distinct groups within the data.

6.1. Scaling

To prepare the data for clustering and dimensionality reduction, numerical features were standardized:

* Standardization: All numerical columns were scaled to have a mean of 0 and a standard deviation of 1.
* Purpose: Standardizing ensured that features with different ranges (e.g., age, percentage of time impacted) were treated equally, preventing larger values from dominating clustering algorithms.
* Impact: The scaled data maintained its structure but became more suitable for distance-based algorithms like KMeans.

6.2. Dimensionality Reduction

High-dimensional data can be challenging to analyze and visualize. UMAP (Uniform Manifold Approximation and Projection) was used for dimensionality reduction:

* Objective: Reduce the data from its original high-dimensional space to a two-dimensional space while preserving important relationships.
* Advantages of UMAP:
  + Retains the global structure of the dataset.
  + Maintains the local density of clusters, making it ideal for visualization.
* Process:
  + The scaled dataset was transformed into two principal components using UMAP.
  + The resulting 2D projection was used to plot clusters visually.
* Impact: Dimensionality reduction simplified data interpretation without losing meaningful cluster characteristics.

6.3. Clustering

After reducing the data, clustering algorithms grouped participants based on their survey responses. KMeans was the chosen algorithm:

* Step 1: Determining Optimal Clusters
  + The "elbow method" was applied using a visualizer to identify the ideal number of clusters (k).
  + The elbow point (where adding more clusters no longer significantly reduces the within-cluster variance) indicated the optimal value for k.
* Step 2: Applying KMeans
  + KMeans partitioned the dataset into k clusters by minimizing the variance within each group.
  + Each participant was assigned to a cluster, representing similar patterns in survey responses.
* Step 3: Evaluating Clusters
  + Cluster quality was assessed using metrics such as:
    - Silhouette Score: Measures how well clusters are separated.
    - Calinski-Harabasz Index: Evaluates the compactness and separation of clusters.
    - Davies-Bouldin Index: Examines the ratio of intra-cluster to inter-cluster distances.

Visualization of Clusters

The clustered data was visualized in the 2D space created by UMAP.

* Cluster Characteristics: Each cluster's main features were analyzed and summarized. For example, clusters could represent employees facing high mental health impacts versus those who did not.
* Graphical Insights: These visualizations highlighted distinct groupings, aiding HR in identifying specific focus areas for targeted mental health interventions.

Impact of Scaling, Reduction, and Clustering

These techniques simplified the complexity of the dataset, enabling meaningful insights into employee groups based on mental health responses. The clusters provided HR with actionable categories, offering a foundation for targeted programs to improve workplace mental health.

Let me know if you'd like any part refined further or visual examples added!

5. References:

https://www.kaggle.com/osmi/mental-health-in-tech-2016