Unveiling Mental Health Patterns in the Tech Industry:
A Comprehensive Clustering and Dimensionality Reduction Analysis

This image was created with the assistance of DALL·E

International University of Applied Sciences

DLBDSMLUSL01
Machine Learning – Unsupervised Learning and Feature Engineering

Case Study
Task 1: Mental Health in Technology-related Jobs

Oskar Wolf
92126079

Table of Contents

1. Introduction

1.1. Background

The mental health of employees in the tech industry has gained significant attention in recent years. High stress, long working hours, and the demanding nature of tech jobs contribute to mental health challenges among professionals in this field. The data for this study is derived from the OSMI Mental Health in Tech Survey 2016, which includes over 1400 responses. The survey aims to measure attitudes towards mental health in the tech workplace and examine the frequency of mental health disorders among tech workers. This dataset provides valuable insights but is limited by its timeframe (2016) and the number of respondents (1433).

1.2. Objective of the Study

This study aims to uncover patterns in mental health within the tech industry using advanced clustering techniques combined with various dimensionality reduction methods. The goal is to identify distinct groups based on their mental health responses and related attributes, providing actionable insights for improving workplace well-being. All code used and referenced in this study is available for download on the GitHub repository.

2. Data Preprocessing

2.1. Data Cleaning

The dataset for this study was available in CSV format as well as a Neo4j JSON file. The preprocessing of the data involved several key steps to ensure it was ready for analysis. All preprocessing code can be accessed from the GitHub repository: Preprocessing Code.

Column Renaming: All headings were renamed using snake_case to ensure consistency and ease of use.

Null Value Handling:

- Null values were found to be associated with specific subsets of data, such as self-employed respondents, current and previous employer related information, certain demographics such as US respondents, and open-ended survey questions for opinion mining. These values were imputed with 'Not Applicable' or appropriate values based on the context.
- Missing values in features related to company size and whether the company was a tech company were due to self-employed respondents. These were imputed to 'Self-Employed'.
- Missing values in the previous employer subset were due to respondents being new to the industry or self-employed. Questions about previous employer benefits and resources were imputed to 'No Previous Employer'.

Country Naming: Countries were renamed to their ISO 3166-1 alpha-3 codes for consistency.

Data Splitting and Encoding:

- Multiple disorder types and work positions features and were split by the '|' character. New columns were created using one-hot encoding to visualize diagnosis and work position demographics.
- Disorder types and work positions were merged categorically with correct naming conventions such as 'y_' for believed current health disorders, 'm_' for possible mental health diagnosis, and 'p_' for professionally diagnosed conditions.

Gender Grouping: Genders were grouped into male, female, and non-binary/other due to the variety of gender types entered in the open text box of the survey. These were categorized into the three groups.

Duplicate Check: No duplicate entries were found in the dataset.

Dropping Features: Features with a high percentage of null values that did not correlate directly to any subset were dropped.

2.2. Encoding Techniques

The encoding of the data was crucial to ensure it was in a format suitable for machine learning models. The encoding process is available from the GitHub repository: Encoding Code.

Steps Taken:

Categorical Data Encoding:

- **Ordinal Categorical Data**: Encoded using LabelEncoder.
- **Nominal Categorical Data**: Encoded using one-hot encoding.

Encode Mappings: Encode mappings were recorded in JSON file for future reference.

Data Type Conversion: All columns were converted to int32 for ordinal encoding or Boolean for nominal encoding.

Generated Outputs:

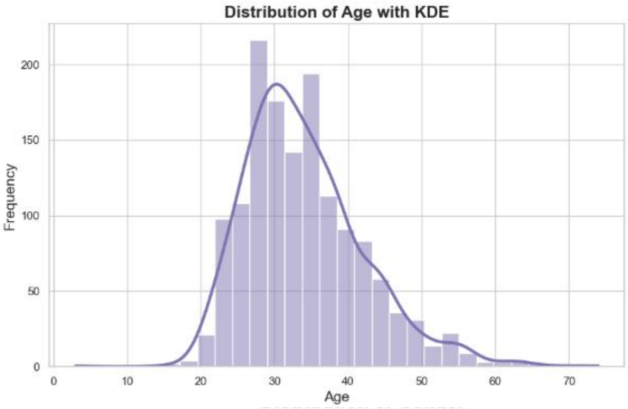
Excel sheets were generated for fully preprocessed and encoded data, along with six various subsets within the preprocessed data for EDA and initial insights. These subsets are available in the GitHub repository: Subsets for EDA.

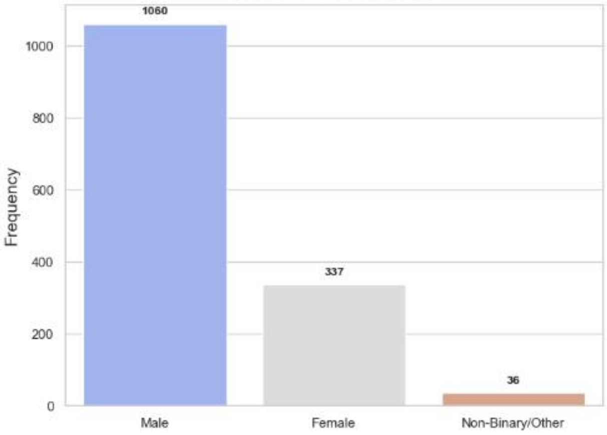
3. Exploratory Data Analysis (EDA)

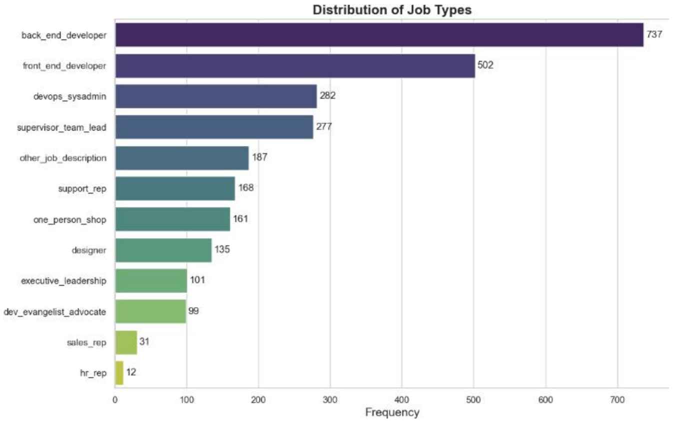
The Global and USA subsets provide a comprehensive view of mental health-related factors on a global scale, including demographic and job-related variables. It includes all pre-processed columns.

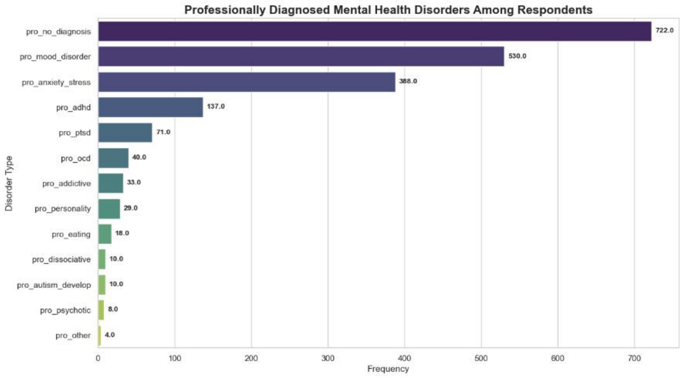
Offers valuable insights into the prevalence and impact of mental health issues across different demographics and job types, informing targeted interventions and policies. Global Subset Code and USA Subset Code

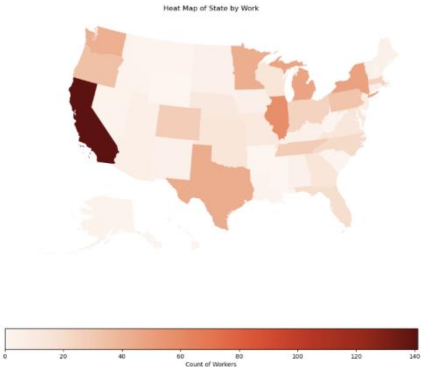
- **Demographics**: The age distribution of respondents centers around the mid-20s to mid-30s, peaking in the late 20s. Gender distribution shows a significant majority of male respondents (1060), followed by females (337) and a smaller group of non-binary/other individuals (36).
- **Job Types**: The most common job types among respondents are back-end developers (737) and front-end developers (502), followed by DevOps/sysadmins (282) and supervisors/team leads (277).
- **Mental Health Disorders**: The most reported professionally diagnosed mental health disorders are mood disorders (530) and anxiety/stress disorders (368), followed by ADHD (137) and PTSD (71). A significant portion of respondents (722) reported no professional diagnosis.
- A world map visualization indicates that the majority of respondents work in the USA, with 59.39% of the total respondents.
- The Jaccard similarity index was used to analyze the relationship between respondents' countries of residence and their countries of work, revealing that 2.05% of US respondents work outside their state of residence and 7.28% of respondents work outside their country of residence.

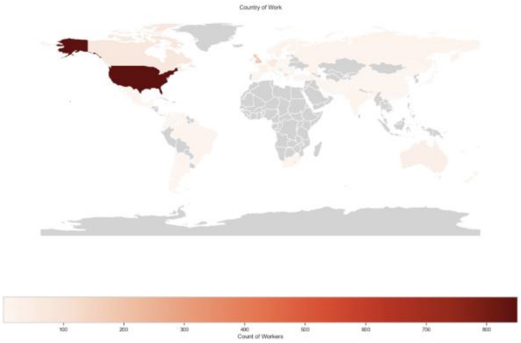












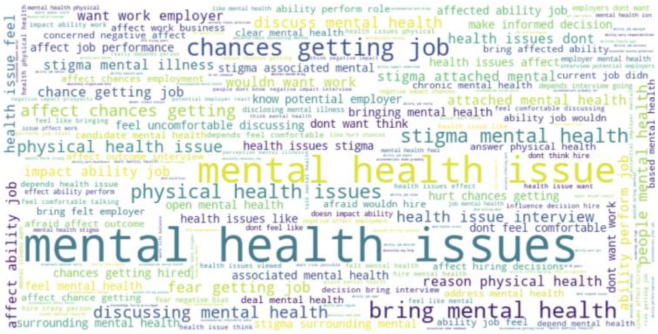
- There are additional visualizations in the GitHub code for insights on self-employed, previous employer views and current employer views but that information is be presented when data is enriched after clustering. Self-Employed Subset Code, Previous Employed Subset Code and Employed Subset Code.

The opinion mining subset focuses on open-ended survey questions related to mental health opinions and experiences. This analysis uses text mining and natural language processing techniques to extract meaningful insights from textual data. Opinion Mining Subset Code.

Summary of Sentiment Analysis for Health Discussions:

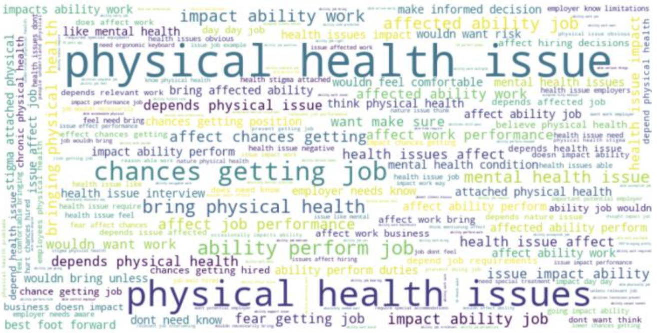
Mental Health:

- Polarity: Neutral overall, reflecting mixed views on discussing mental health in interviews.
- Subjectivity: Moderate, with responses mixing personal opinions and factual information.



Physical Health:

- Polarity: Slightly positive overall, indicating more willingness to discuss physical health.
- Subjectivity: Moderately high, mixing personal opinions and factual statements.



Implications:

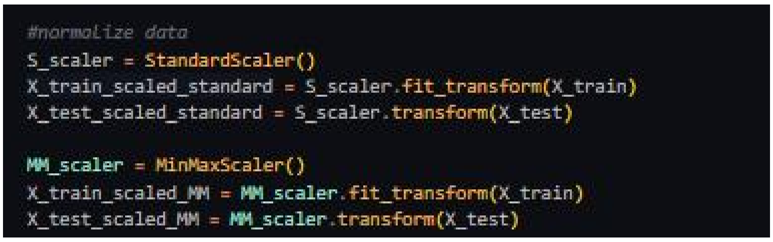
The overall comparison shows that physical health discussions are generally more positive, implying that candidates feel safer discussing physical health conditions than mental health conditions. The negative sentiment around mental health discussions points to ongoing stigmas and fears of bias. Both topics exhibit moderate subjectivity, emphasizing the personal nature of health discussions. However, the slightly higher subjectivity in physical health discussions indicates a greater variability in individual perceptions and reactions. The analysis reflects broader societal views where physical health issues are less stigmatized than mental health issues, influencing how openly individuals discuss these topics during job interviews. Organizations should promote inclusivity and support to reduce negative perceptions and encourage openness in health-related discussions.

4. Feature Importance and Selection Methods

In the process of determining which features are most relevant for analysis, various techniques were employed. These techniques help in understanding the data better and ensuring that only the most informative features are used in the modeling process.

Standard Scaler

- Purpose: Provided specific insights by focusing on key features, which helps in better cluster separation and compactness.



```
#normalize data
S_scaler = StandardScaler()
X_train_scaled_standard = S_scaler.fit_transform(X_train)
X_test_scaled_standard = S_scaler.transform(X_test)

MM_scaler = MinMaxScaler()
X_train_scaled_MM = MM_scaler.fit_transform(X_train)
X_test_scaled_MM = MM_scaler.transform(X_test)
```

- Analysis: The selected features emphasize personal mental health history and experiences with previous employers.
- Advice: This aligns well with the goal of identifying key factors related to mental health in the tech industry. Consider the top features for deeper analysis.

MinMax Scaler

- Purpose: Scales features to a specific range, which is beneficial for algorithms sensitive to the magnitude of data, like KMeans and DBSCAN, leading to improved clustering results.
- Analysis: This includes demographic information and organizational factors.
- Advice: Ensure demographic factors are used appropriately to avoid biased outcomes.

The choice was made to go with Standard Scaler given the nature of the data.

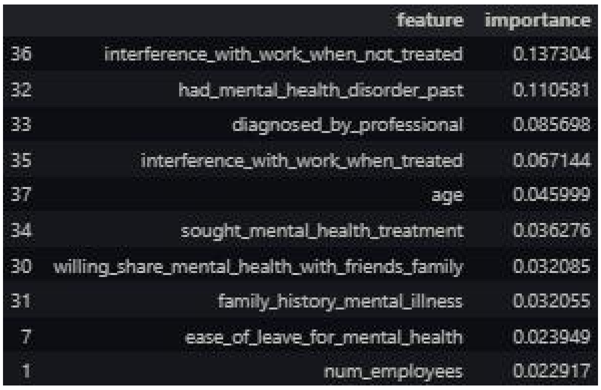
4.1 Supervised Feature Selection

The model was trained on 35 Features based off PCA and selected using the best results from the following methods:

RandomForest:

Provides feature importance scores, indicating the contribution of each feature to the prediction. Rank features based on their importance scores.

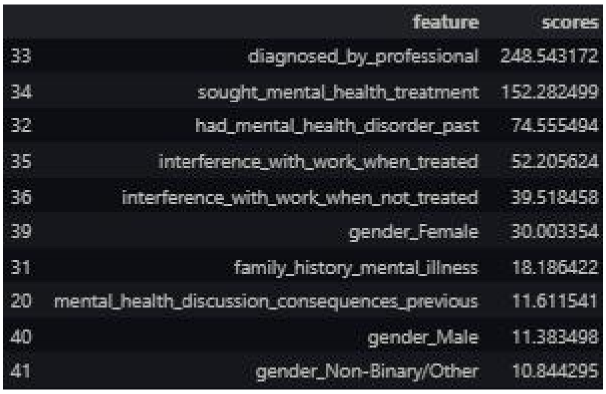
Choose a threshold for selecting top features. This can be done by looking at the cumulative importance or selecting the top N features.



	feature	importance
36	interference_with_work_when_not_treated	0.137304
32	had_mental_health_disorder_past	0.110581
33	diagnosed_by_professional	0.085698
35	interference_with_work_when_treated	0.067144
37	age	0.045999
34	sought_mental_health_treatment	0.036276
30	willing_share_mental_health_with_friends_family	0.032085
31	family_history_mental_illness	0.032055
7	ease_of_leave_for_mental_health	0.023949
1	num_employees	0.022917

SelectKBest (chi2):

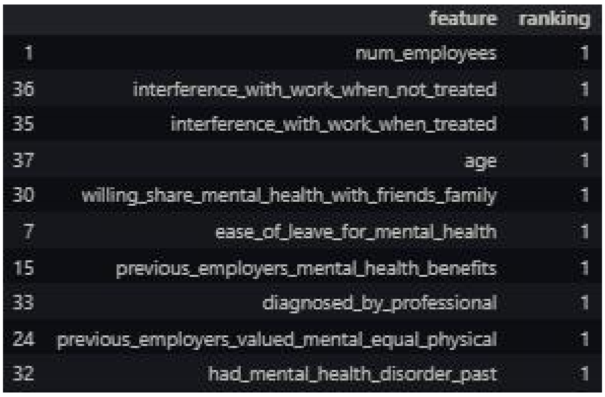
Uses statistical tests to score features based on their relevance to the target variable. Rank features based on their scores, Choose the number of top features (k) based on their scores or a desired variance threshold.



	feature	scores
33	diagnosed_by_professional	248.543172
34	sought_mental_health_treatment	152.282499
32	had_mental_health_disorder_past	74.555494
35	interference_with_work_when_treated	52.205624
36	interference_with_work_when_not_treated	39.518458
39	gender_Female	30.003354
31	family_history_mental_illness	18.186422
20	mental_health_discussion_consequences_previous	11.611541
40	gender_Male	11.383498
41	gender_Non-Binary/Other	10.844295

RFE (Recursive Feature Elimination):

Recursively eliminates the least important features based on the model's coefficients or feature importance. Features are ranked, with a lower ranking indicating higher importance, Select features based on their ranking, often the top N features.

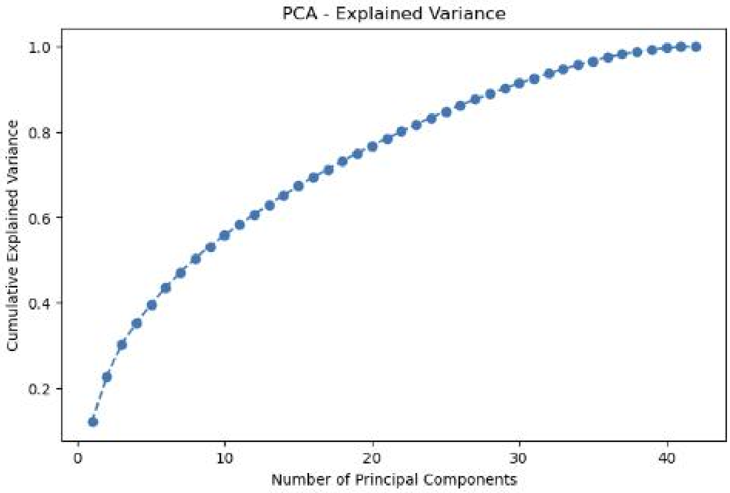


	feature	ranking
1	num_employees	1
36	interference_with_work_when_not_treated	1
35	interference_with_work_when_treated	1
37	age	1
30	willing_share_mental_health_with_friends_family	1
7	ease_of_leave_for_mental_health	1
15	previous_employers_mental_health_benefits	1
33	diagnosed_by_professional	1
24	previous_employers_valued_mental_equal_physical	1
32	had_mental_health_disorder_past	1

PCA (Principal Component Analysis):

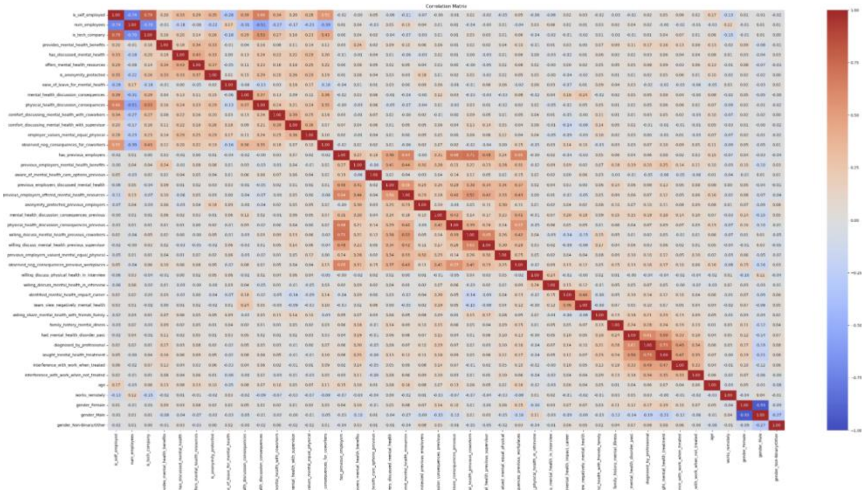
Transforms the features into a set of linearly uncorrelated components. The number of components to select depends on the cumulative explained variance.

Plot the cumulative explained variance,Choose the number of components that explain a sufficient amount of variance.



Correlation Matrix:

The correlation matrix helps identify the relationships between features, indicating which features are strongly correlated. Highly correlated features may contain redundant information and can be considered for removal to reduce multicollinearity.

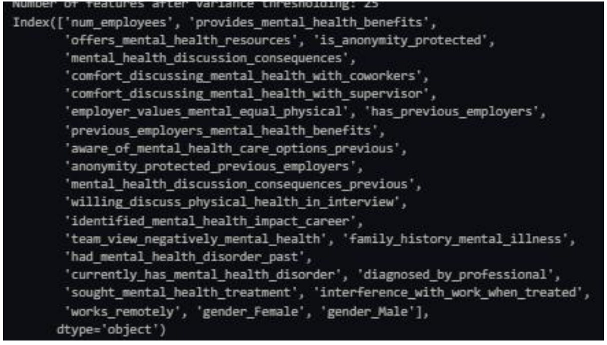


4.2 Unsupervised Feature Selection

The features that are selected below are a combination of all feature selection techniques:

Variance Threshold:

An initial filter for removing features with very low variance, which are likely uninformative. Retained 25 features, effectively reducing the feature space

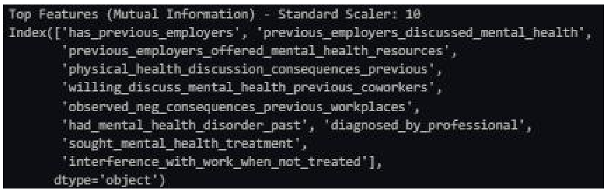


```
Number of features after variance thresholding: 25
Index(['num_employees', 'provides_mental_health_benefits',
       'offers_mental_health_resources', 'is_anonymity_protected',
       'mental_health_discussion_consequences',
       'comfort_discussing_mental_health_with_coworkers',
       'comfort_discussing_mental_health_with_supervisor',
       'employer_values_mental_equal_physical', 'has_previous_employers',
       'previous_employers_mental_health_benefits',
       'aware_of_mental_health_care_options_previous',
       'anonymity_protected_previous_employers',
       'mental_health_discussion_consequences_previous',
       'willing_discuss_physical_health_in_interview',
       'identified_mental_health_impact_career',
       'team_view_negatively_mental_health', 'family_history_mental_illness',
       'had_mental_health_disorder_past',
       'currently_has_mental_health_disorder', 'diagnosed_by_professional',
       'sought_mental_health_treatment', 'interference_with_work_when_treated',
       'works_remotely', 'gender_Female', 'gender_Male'],
      dtype='object')
```

Mutual Information:

Measures the dependency between each feature and the target variable.

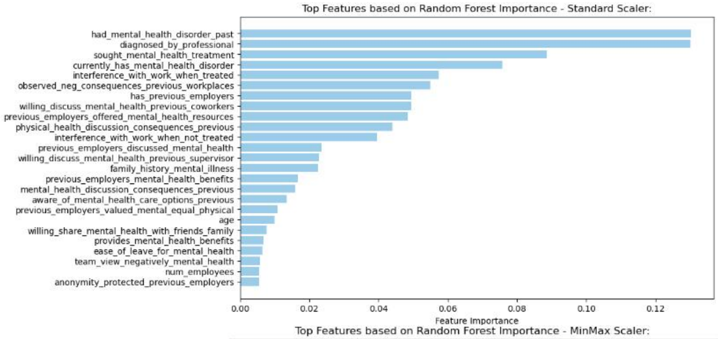
Provided 10 features.



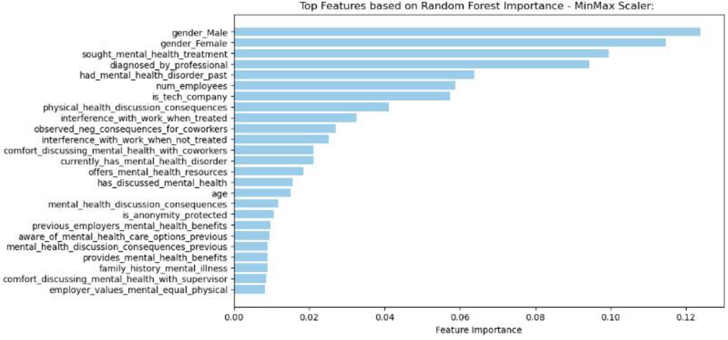
```
Top Features (Mutual Information) - Standard Scaler: 10
Index(['has_previous_employers', 'previous_employers_discussed_mental_health',
       'previous_employers_offered_mental_health_resources',
       'physical_health_discussion_consequences_previous',
       'willing_discuss_mental_health_previous_coworkers',
       'observed_neg_consequences_previous_workplaces',
       'had_mental_health_disorder_past', 'diagnosed_by_professional',
       'sought_mental_health_treatment',
       'interference_with_work_when_not_treated'],
      dtype='object')
```

RandomForest Feature Importance:

- ➢ Standard Scaler Analysis: Emphasizes personal mental health history and experiences with previous employers.



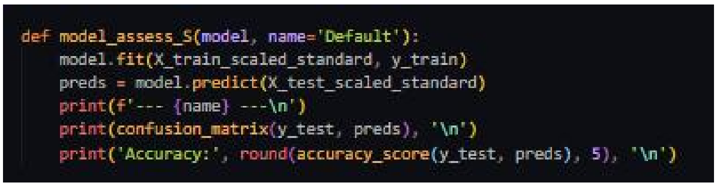
- ➢ MinMax Scaler Analysis: Includes demographic information and organizational factors. Ensured demographic factors are used appropriately to avoid biased outcomes.



Code For Supervised Feature Selection and Modelling: Supervised Model Code

5. Supervised Learning Model

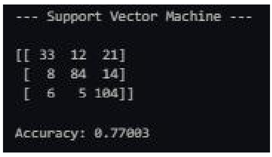
5.1. Supervised Modeling Algorithms



```
def model_assess_S(model, name='Default'):
    model.fit(X_train_scaled_standard, y_train)
    preds = model.predict(X_test_scaled_standard)
    print(f'--- {name} ---\n')
    print(confusion_matrix(y_test, preds), '\n')
    print('Accuracy:', round(accuracy_score(y_test, preds), 5), '\n')
```

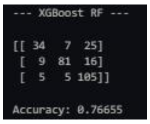
In this section, various supervised learning algorithms were tested to classify mental health-related data. The performance of each model was assessed using a custom model_assess function, which evaluated models based on their confusion matrix and accuracy score. Below is a summary of the top algorithms tested and their key characteristics:

- **SVM** is a linear classifier that finds the hyperplane that best separates different classes. It is effective for high-dimensional spaces and works well for both linear and non-linear data.



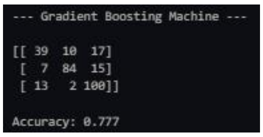
```
--- Support Vector Machine ---

[[ 33  12  21]
 [  8  84  14]
 [  6   5 104]]

Accuracy: 0.77003
```

- **XGBRFClassifier** combines gradient boosting and random forest techniques to improve performance. It leverages the strengths of both methods.



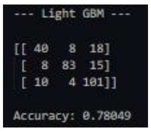
```
--- XGBoost RF ---

[[ 34   7  25]
 [  9  81  16]
 [  5   5 105]]

Accuracy: 0.76655
```

- **GBM** is an ensemble technique that builds trees sequentially, with each tree trying to correct the errors of the previous one. It is powerful but can be slow and prone to overfitting.



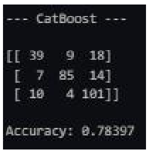
```
--- Gradient Boosting Machine ---

[[ 39  10  17]
 [  7  84  15]
 [ 13   2 100]]

Accuracy: 0.777
```

- **LightGBM** is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithms. It is designed for efficient handling of large datasets.



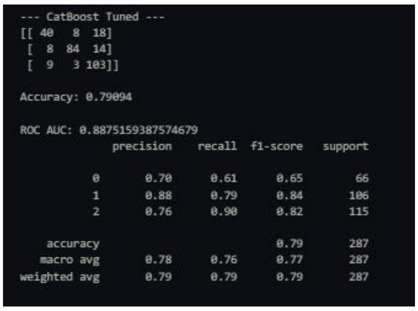
```
--- Light GBM ---

[[ 40   8  18]
 [  8  83  15]
 [ 10   4 101]]

Accuracy: 0.78049
```

- **CatBoost** is a gradient boosting algorithm that handles categorical features automatically and efficiently. It reduces the need for extensive preprocessing.



```
--- CatBoost ---

[[ 39   9  18]
 [  7  85  14]
 [ 10   4 101]]

Accuracy: 0.78397
```

- There were various other tested methods such as Naive Bayes, Stochastic Gradient Descent, KNN, Decision Trees, Random Forest, Neural Networks, Logistic Regression, XG Boost, AdaBoost, Extra Trees Classifier, LDA , QDA, GMM, Ridge Classifier, Lasso Regression and Elastic Net.

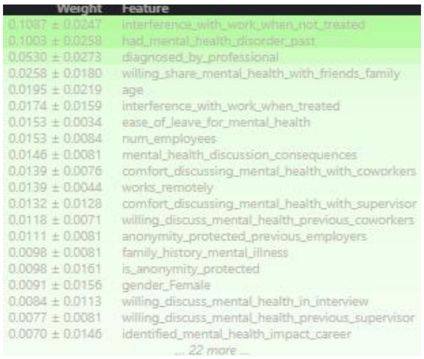


```
--- CatBoost Tuned ---
[[ 40   8  18]
 [  8  84  14]
 [  9   3 103]]

Accuracy: 0.79094

ROC AUC: 0.8875159387574679
              precision    recall  f1-score   support

           0       0.70      0.61      0.65        66
           1       0.88      0.79      0.84       106
           2       0.76      0.90      0.82       115

    accuracy                           0.79       287
   macro avg       0.78      0.76      0.77       287
weighted avg       0.79      0.79      0.79       287
```

5.2. Parameter Tuning and Final Model

The CatBoost model demonstrates good performance in predicting the mental health disorder status among tech professionals. The high ROC AUC score indicates strong discriminatory power, while the balanced precision, recall, and F1-scores across classes suggest the model handles each class well.

However, the model performs slightly better in identifying individuals with a current mental health disorder (Class 1) and those who may have a disorder (Class 2) compared to those with no current disorder (Class 0).



Weight	Feature
0.1087 ± 0.0247	interference_with_work_when_not_treated
0.1003 ± 0.0258	had_mental_health_disorder_past
0.0530 ± 0.0273	diagnosed_by_professional
0.0258 ± 0.0180	willing_share_mental_health_with_friends_family
0.0195 ± 0.0219	age
0.0174 ± 0.0159	interference_with_work_when_treated
0.0153 ± 0.0034	ease_of_leave_for_mental_health
0.0153 ± 0.0084	num_employees
0.0146 ± 0.0081	mental_health_discussion_consequences
0.0139 ± 0.0076	comfort_discussing_mental_health_with_coworkers
0.0139 ± 0.0044	works_remotely
0.0132 ± 0.0128	comfort_discussing_mental_health_with_supervisor
0.0118 ± 0.0071	willing_discuss_mental_health_previous_coworkers
0.0111 ± 0.0081	anonymity_protected_previous_employers
0.0098 ± 0.0081	family_history_mental_illness
0.0098 ± 0.0161	is_anonymity_protected
0.0091 ± 0.0156	gender_Female
0.0084 ± 0.0113	willing_discuss_mental_health_in_interview
0.0077 ± 0.0081	willing_discuss_mental_health_previous_supervisor
0.0070 ± 0.0146	identified_mental_health_impact_career

… 22 more …

Code For Supervised Feature Selection and Modelling: Clustering Model

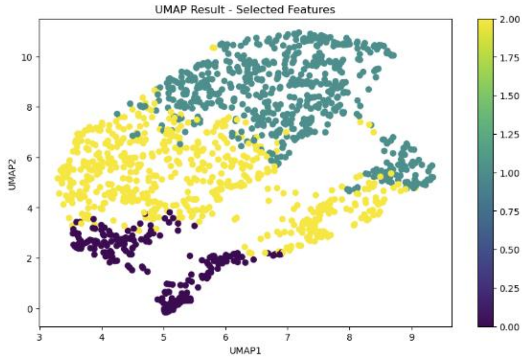
6. Unsupervised Learning and Clustering

For the most optimal clustering model a selection of dimensionality reduction techniques were used with a variety of clustering algorithms to determine the best fit.

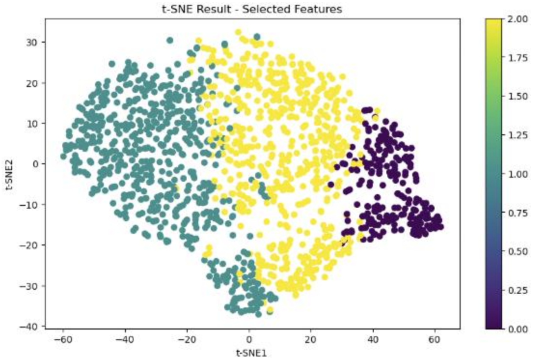
6.1. Dimensionality Reduction Techniques

Other methods were also tested, some with good metrics, however they gave insufficient results to represent real life information such as Isomap, Local Linear Embedding, Principal Component Analysis, Independent Component Analysis, Multidimensional Scaling, Multiple Correspondence Analysis but most reliable and prominent were:

- **UMAP (Uniform Manifold Approximation and Projection)**: Preserves both global and local structure of the data while creating a low-dimensional representation that maintains the topological structure of the original space. The scatter plot shows clear and distinct cluster separation, with clusters well-distributed across the UMAP1 and UMAP2 axes.



- **t-SNE (t-Distributed Stochastic Neighbor Embedding)**: Focuses on preserving local similarities, creating a map where similar data points are modeled by nearby points and dissimilar points are modeled by distant points.The scatter plot shows distinct clusters with clear separation along the t-SNE axes.



6.2. Clustering Techniques

Many clustering techniques were tested such as Mean Shift, Affinity Propagation, OPTICS, Gaussian Mixture Models, DBSCAN, Spectral Clustering but the most reliable and representable of real world data were:

- **KMeans Clustering:** Partitions data into k clusters by minimizing the variance within each cluster.
- **Agglomerative Clustering:** A hierarchical clustering method that builds nested clusters by successively merging or splitting them based on a distance criterion.

6.3. Evaluation Metrics

Models were accessed using:

Silhouette Score: Measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to +1, where a higher score indicates better-defined clusters.

Davies-Bouldin Index: Measures the average similarity ratio of each cluster with the cluster that is most similar to it. Lower values indicate better separation between clusters.

Calinski-Harabasz Index: Measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion. Higher values indicate better-defined clusters.

K Elbow Criterion: This method helps in determining the optimal number of clusters (K) in a dataset by plotting the explained variation as a function of the number of clusters. The "elbow point" on the plot indicates the number of clusters where the addition of another cluster does not significantly improve the clustering performance.

7. Results and Discussion

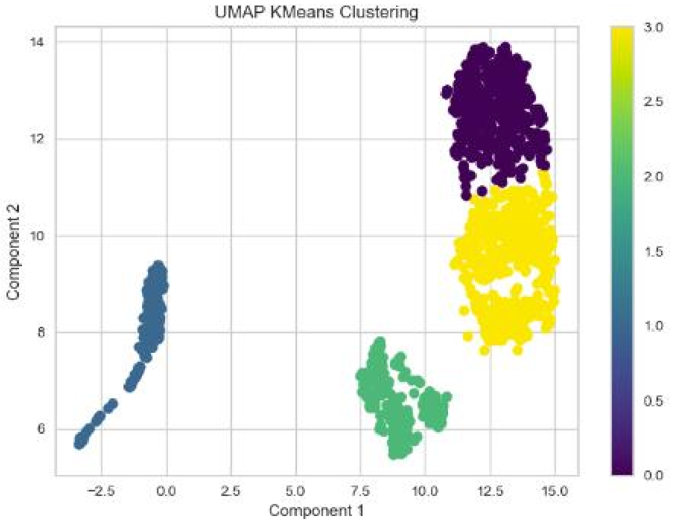
7.1. Best Combination of Techniques

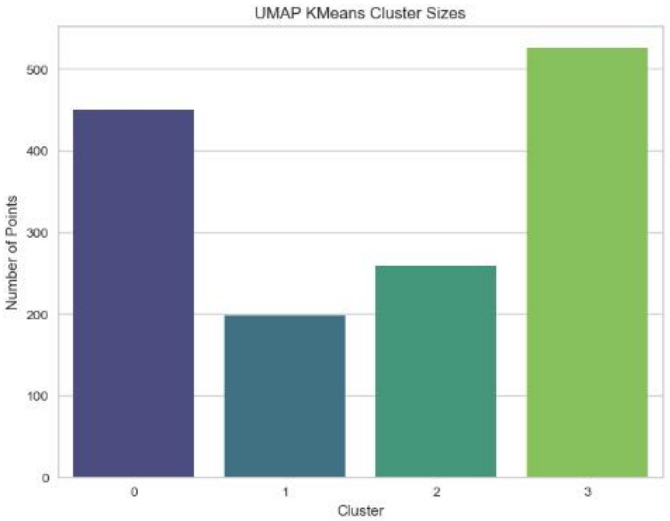
Considering all possible combinations of techniques, metrics and how well clusters represent real world data, **UMAP KMeans** stands out as the best overall combination.

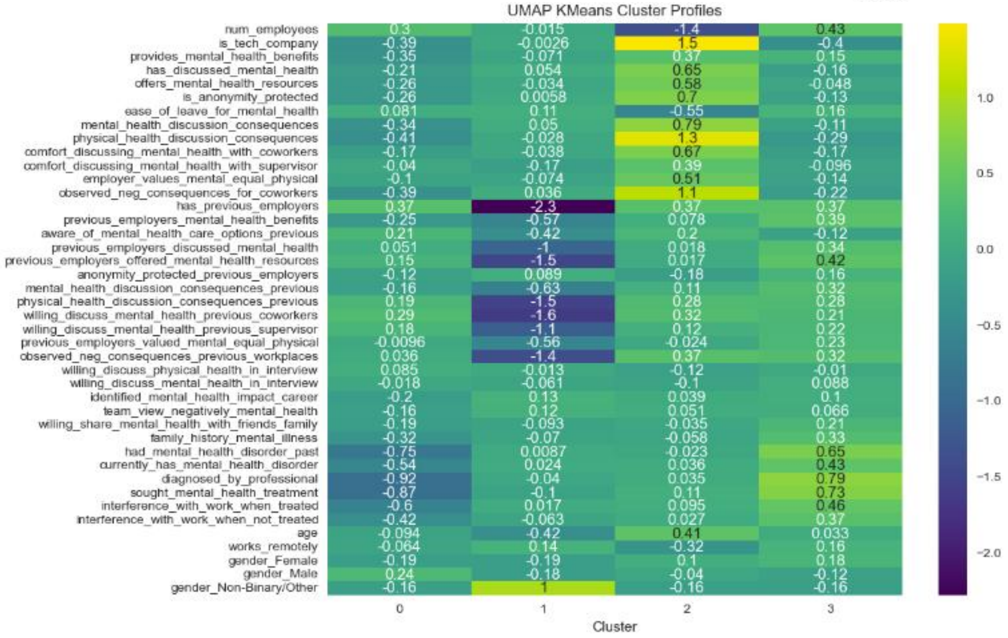
Silhouette Score: 0.6085922718048096

Davies-Bouldin Index: 0.520276936698626

Calinski-Harabasz Index: 8557.146559995392







7.2. Insights from Cluster Profiles

Cluster 0:

- **Mental Health: Least likely to suffer from a mental health disorder.**
- Employment History: Have had previous employers.
- Family History: Least likely to have a family history of mental health issues.
- Company Benefits: Likely to have access to mental health benefits.
- Work Environment: More likely to feel comfortable discussing mental health with coworkers and supervisors.

Cluster 1:

- **Demographics: Most likely to be younger and on their first job. Contains all non-binary gender individuals**.
- Mental Health: New in the work environment. Third most likely to suffer from a mental health disorder.
- Work Interference: Less likely to experience work interference when mental health issues are treated.
- Support Systems: Less likely to have sought mental health treatment or to be diagnosed by a professional.

Cluster 2:

- **Employment History: Most likely to be older and self-employed.**
- Mental Health: Can have mental health issues due to the stress of running a business. Second most likely to suffer from a mental health disorder.
- Work Interference: Likely to experience work interference both when treated and untreated.
- Support Systems: More likely to have sought mental health treatment.

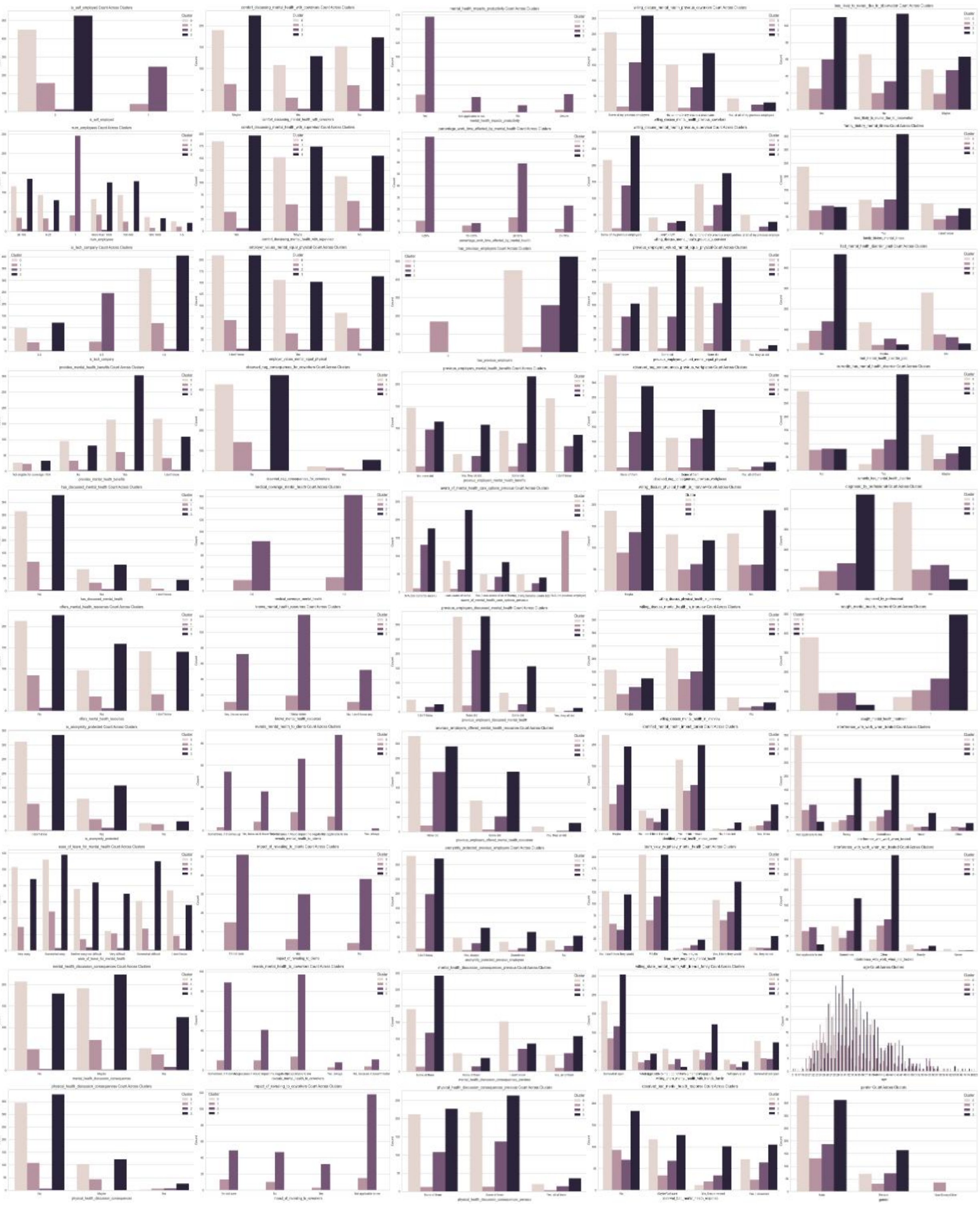
Cluster 3:

- **Mental Health: Most likely to suffer from a mental health disorder. Diagnosed by a professional and had mental health disorders in the past.**
- Family History: Most likely to have a family history of mental health issues.
- Work Interference: Work interference often when not treated.
- Employment History: Have had previous employers.
- Company Benefits: More likely to have access to mental health benefits and feel comfortable discussing mental health at work.

General Observations:

- Company Size: Employees in larger companies are more likely to have access to mental health benefits.
- Anonymity Protection: Those who feel their anonymity is protected are more likely to discuss mental health issues.

- Remote Work: Distribution of remote work varies across clusters, with some clusters showing a higher likelihood of working remotely.



8. References

Websites:

- Kaggle. (2016). OSMI Mental Health in Tech Survey 2016. Retrieved from https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016/data
- Creative Commons. (n.d.). Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Retrieved from https://creativecommons.org/licenses/by-sa/4.0/deed.en
- GitHub. (2023). Subsets for EDA. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/tree/main/Subsets
- GitHub. (2023). ML-EDA-Mental-Health-In-Tech. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/tree/main

Libraries Used:

- ‘seaborn’ 0.13.2
- ‘matplotlib’ 3.8.0
- ‘pandas’ 2.1.4
- ‘numpy’ 1.26.3
- ‘geopandas’ 0.14.4
- ‘textblob’ 0.15.3
- ‘wordcloud’ 1.9.3
- ‘sklearn-learn’ 1.2.2
- ‘xgboost’ 2.0.3
- ‘lightgbm’ 4.3.0
- ‘catboost’ 1.2.5
- ‘eli5’ 0.13.0

Code References:

- GitHub. (2023). Preprocessing Code. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/Preprocessing/preprocessing.ipynb
- GitHub. (2023). Encoding Code. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/Preprocessing/encode_data.ipynb
- GitHub. (2023). Encoded Mappings. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/Preprocessing/encoded_mappings.json
- GitHub. (2023). EDA Global. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/EDA/EDA-Global.ipynb
- GitHub. (2023). EDA USA. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/EDA/EDA-USA.ipynb

- GitHub. (2023). EDA Self-Employed. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/EDA/EDA-Self-Employed.ipynb
- GitHub. (2023). EDA Previous Employed. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/EDA/EDA-Previous-Employed.ipynb
- GitHub. (2023). EDA Opinion Mine. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/EDA/EDA-Opinion-Mine.ipynb
- GitHub. (2023). EDA Employed. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/EDA/EDA-Employed.ipynb
- GitHub. (2023). Supervised Model. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/ML/supervised_model.ipynb
- GitHub. (2023). Clustering Model. Retrieved from https://github.com/oskar-wolf/ML-EDA-Mental-Health-In-Tech/blob/main/ML/clustering_model.ipynb