**DATA PREPARATION & ANALYSIS (DPA) PROJECT REPORT**

**Course: CSP 571**

**MENTAL HEALTH IN TECHNOLOGY**

**Submitted by:**

Ayesha Saif (A20562668)

Azha Manzoor (A20557552)

Dhruval Patel (A20549909)

Contents

[**Abstract** 2](#_Toc183619215)

[**Introduction** 3](#_Toc183619216)

[**Aim & Objectives** 5](#_Toc183619217)

[**Dataset Description** 7](#_Toc183619218)

[**Exploratory Data Analysis (EDA)** 9](#_Toc183619219)

[**Correlation Analysis** 11](#_Toc183619220)

[**Dimensionality Reduction Strategies** 13](#_Toc183619221)

[**Unsupervised Learning Techniques** 15](#_Toc183619222)

[**Cross-Validation Strategy** 16](#_Toc183619223)

[**Model Training & Performance Analysis** 17](#_Toc183619224)

[**Improvements & Results** 18](#_Toc183619225)

[**Recommendations** 19](#_Toc183619226)

### **Abstract**

The research delves into the critical issue of mental health among technology practitioners, specifically examining the impact of remote work and familial history of mental illness. By conducting a comprehensive exploratory data analysis (EDA) on a substantial dataset, we aimed to uncover significant patterns and correlations. Key findings include the influence of workplace benefits on mental health outcomes.

To further investigate the factors associated with help-seeking behavior, we employed a variety of machine learning techniques, including logistic regression, random forest, and clustering algorithms. Our results highlight the potential for targeted interventions to improve mental health support within the tech industry. Among the models evaluated, random forest achieved the highest validation accuracy of 83.33%, demonstrating its effectiveness in predicting help-seeking behavior.

The insights gained from this study have profound implications for tech employers, employees, and mental health professionals. By understanding the complex interplay of factors contributing to mental health challenges, we can develop evidence-based strategies to promote well-being and reduce stigma. Future research directions may involve exploring the long-term effects of remote work, identifying additional risk factors, and evaluating the effectiveness of specific interventions.

### **Introduction**

The technology industry, renowned for its innovation and rapid pace, has simultaneously become a breeding ground for mental health challenges. Prolonged work hours, constant pressure to stay updated with the latest trends, and the demanding nature of the work environment have taken a significant toll on the mental well-being of tech professionals. While the importance of mental health has gained recognition in recent years, the stigma surrounding mental illness persists, hindering individuals from seeking the necessary support.

Remote work, a prevalent trend in the tech industry, has introduced new complexities to mental health. While it offers flexibility, it can also lead to feelings of isolation, blurred work-life boundaries, and increased stress. Additionally, a family history of mental illness can predispose individuals to mental health issues, making them more vulnerable to the challenges of the tech industry.

To address these concerns, this study leverages a large dataset from a mental health survey to investigate the treatment patterns among tech professionals. By analyzing numerous factors, including remote work, family history of mental illness, and workplace culture, we aim to identify key determinants of help-seeking behavior. Through exploratory data analysis (EDA), we delve deeper into the relationship between mental health and factors such as workplace conversations about mental health and employer support.

To predict help-seeking behavior, we employ a combination of machine learning techniques, including logistic regression, random forest, and clustering algorithms. These models enable us to identify patterns and trends that can inform targeted interventions. By understanding the factors that influence help-seeking behavior, we can develop strategies to encourage individuals to seek the support they need.

The findings of this study have significant implications for employers, policymakers, and mental health professionals. By implementing workplace initiatives that promote mental well-being, such as flexible work arrangements, mental health awareness programs, and access to mental health resources, employers can create a healthier and more productive work environment. Policymakers can leverage these insights to develop evidence-based policies that address the unique mental health needs of tech professionals. Mental health professionals can tailor their interventions to the specific challenges faced by individuals in the tech industry, providing effective and compassionate care.

### **Aim & Objectives**

To analyze mental health treatment trends among technology professionals, identify key factors influencing treatment-seeking behavior, and develop predictive models to support targeted interventions, improving mental well-being and fostering a supportive workplace environment.

**Objectives:**

1. **Data Exploration**:
   * Conduct a thorough analysis of the dataset to understand the relationships between features, their impact on mental health treatment, and evaluate the assumptions of feature independence.
   * Identify key correlations and patterns that influence treatment-seeking behavior.
2. **Visualization and Dimensionality Reduction**:
   * Employ various visualization techniques, including correlation plots, PCA, UMAP, and t-SNE, to uncover insights and assess data structure and clustering.
3. **Unsupervised Learning Analysis**:
   * Apply unsupervised learning techniques such as clustering to explore hidden patterns and groupings within the data.
4. **Model Development and Evaluation**:
   * Train a baseline predictive model for mental health treatment outcomes.
   * Utilize a stratified cross-validation strategy to assess model performance and ensure robust evaluation.
5. **Optimization and Performance Improvement**:
   * Conduct experiments to enhance model performance through feature selection, regularization, and tuning model complexity.
   * Explore advanced modeling approaches like Random Forest and boosting techniques to improve prediction accuracy.
6. **Actionable Insights**:
   * Provide recommendations for employers, policymakers, and mental health professionals based on findings to foster mental health awareness and improve access to care in the tech industry.

### **Dataset Description**

The dataset used in this project, sourced from a mental health survey on Kaggle, provides comprehensive insights into mental health challenges faced by professionals in the technology industry. It consists of 27 features covering demographic, personal, and workplace factors, with the primary goal of understanding the predictors of mental health treatment-seeking behavior. Below is a detailed breakdown of the dataset:

1. **Demographic Information**:
   * **Age**: Represents the age of respondents, allowing for analysis of mental health patterns across different age groups.
   * **Gender**: Includes self-reported gender, used to explore gender-based disparities in mental health trends.
2. **Family and Medical History**:
   * **Family History**: Indicates whether respondents have a family history of mental illness, a significant factor influencing mental health treatment.
   * **Personal Medical History**: Covers past and present mental health conditions.
3. **Workplace Factors**:
   * **Remote Work**: Identifies if the respondent works remotely, exploring its impact on mental health.
   * **Company Benefits**: Includes the availability of mental health benefits, wellness programs, and leave policies in the workplace.
   * **Workplace Culture**: Explores factors such as supervisor support and comfort discussing mental health at work.
4. **Mental Health Outcomes**:
   * **Treatment-Seeking Behavior**: The target variable indicates whether the respondent has sought treatment for a mental health condition.
5. **Survey-Specific Details**:
   * Includes data on anonymity in mental health discussions, geographic location, and accessibility of mental health resources.

**Data Characteristics**

* **Size**: The dataset contains several hundred responses, providing a statistically significant sample for analysis. A larger sample size enhances the reliability of the findings.
* **Categorical and Numerical Data**: The dataset includes a mix of categorical (e.g., gender, benefits) and numerical features (e.g., age), necessitating preprocessing steps like encoding and normalization to ensure the data is ready for machine learning models.
* **Missing Values**: Some features contain missing values, which are addressed during data cleaning to maintain the integrity of the analysis. Handling missing data is crucial for accurate and reliable results.
* **Balanced vs. Imbalanced Features**: The target variable may exhibit class imbalance, requiring techniques like stratified sampling to prevent biased modeling outcomes. Ensuring balanced classes helps in developing models that perform well across all categories.

This dataset offers a rich foundation for analyzing mental health treatment in the tech industry, enabling the identification of key factors and actionable insights through advanced exploratory and predictive modeling techniques.

### **Exploratory Data Analysis (EDA)**

The dataset underwent thorough analysis to extract meaningful insights into mental health trends among technology professionals:

* **Outlier Detection**: Anomalies in age and workplace factors were identified and either excluded or transformed to prevent skewing the results.
* **Feature Transformation**: Numerical features were normalized to ensure compatibility with machine learning models and enhance their performance.
* **Category Distribution**: The distribution of categorical data was assessed, highlighting imbalances in gender representation and treatment-seeking behavior.
* **Target Variable Analysis**: The distribution of treatment-seeking outcomes was explored, revealing potential biases related to workplace benefits and family history.

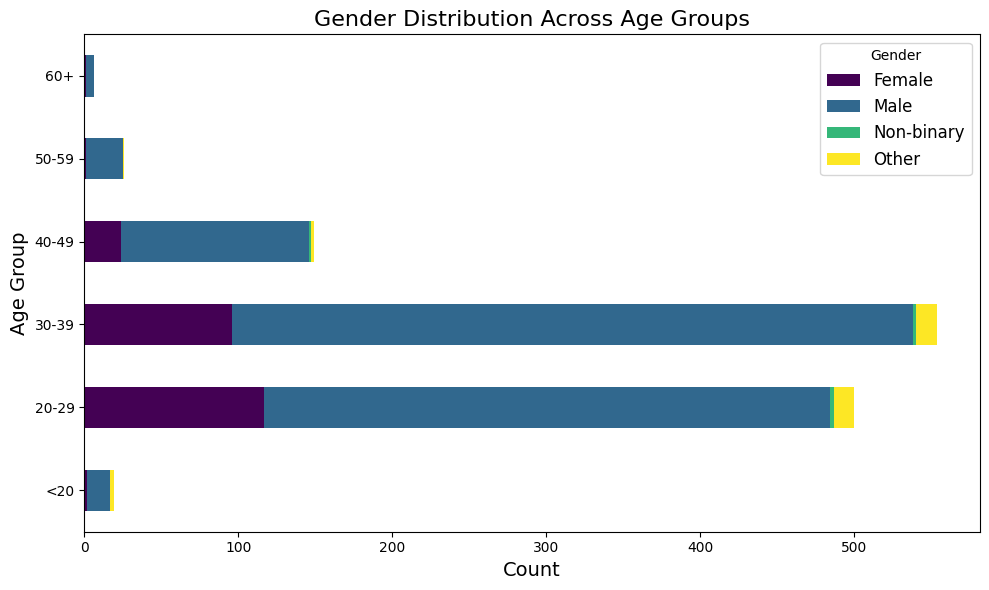


Figure : Gender & Age Distribution

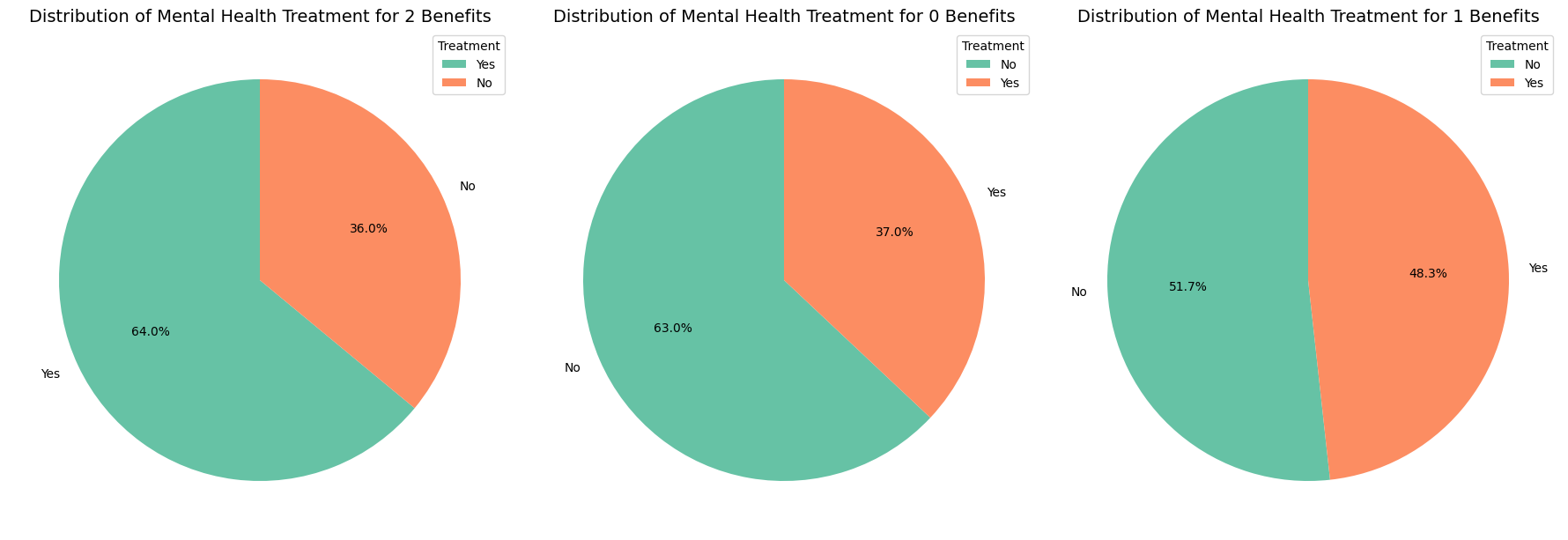
****

Figure : Company Benefits Impact on Treatment

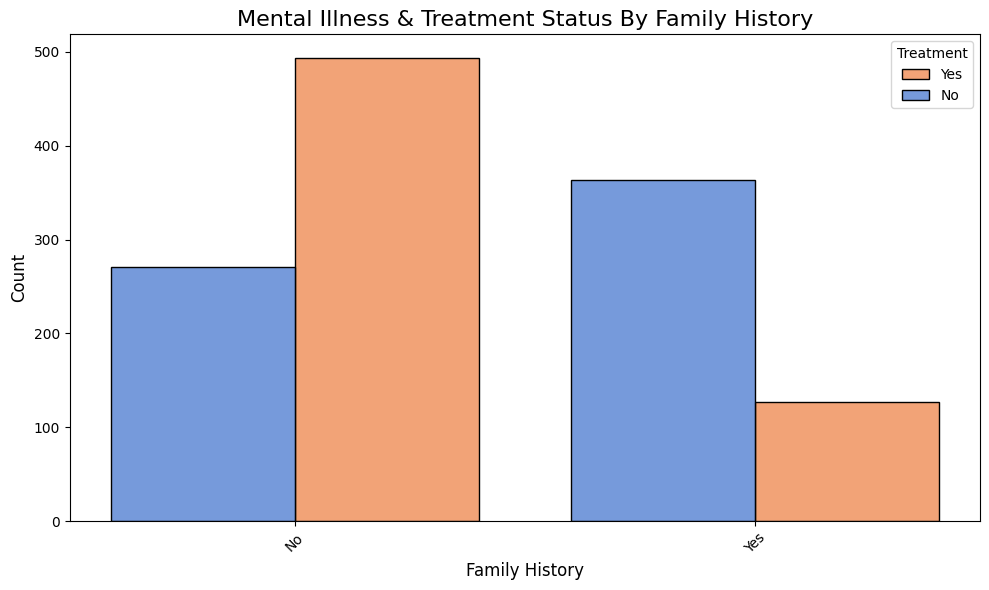
****

Figure : Mental Health & Treatment by Family History

### **Correlation Analysis**

**Strong Positive Correlation:**

* **Mental & Physical Health Consequences:** A robust connection exists between mental and physical health. Individuals experiencing mental health challenges are more likely to suffer from physical health issues as well.
* **Leave Taken:** A strong association was found between the amount of leave taken and both mental and physical health consequences. This suggests that individuals who take more leave may be experiencing higher levels of stress or illness.
* **Support Network:** A positive correlation emerged between the perceived level of support from coworkers and supervisors and mental health. A strong support network can significantly contribute to positive mental health outcomes.
* **Health Discussions:** Discussions about mental health often coincide with discussions about physical health, indicating a growing awareness of the interconnectedness of these two aspects of well-being.

**Moderate Positive Correlation:**

* **Care Options & Benefits:** Access to mental health care options and benefits was positively correlated, suggesting that companies offering these resources are more likely to prioritize employee well-being.
* **Wellness Programs:** Companies with wellness programs often provided mental health care options as part of their overall employee benefits package.
* **Anonymity & Leave:** Anonymity in seeking mental health support was associated with higher rates of leave-taking, possibly due to reduced stigma and increased comfort in accessing care.
* **Health Perception:** Individuals who recognized the link between mental and physical health were more likely to engage in discussions about mental health, indicating a greater awareness of the importance of holistic well-being.

**Negative Correlation:**

* **Year vs. Month/Day:** A negative correlation was observed between the year and calendar attributes (month and day). This is expected as it reflects the passage of time and the changing seasons.
* **Mental-Physical Link & Observations:** A stronger belief in the connection between mental and physical health was associated with a reduced perception of negative consequences. This suggests that individuals who understand this link may be better equipped to manage their overall health and well-being.

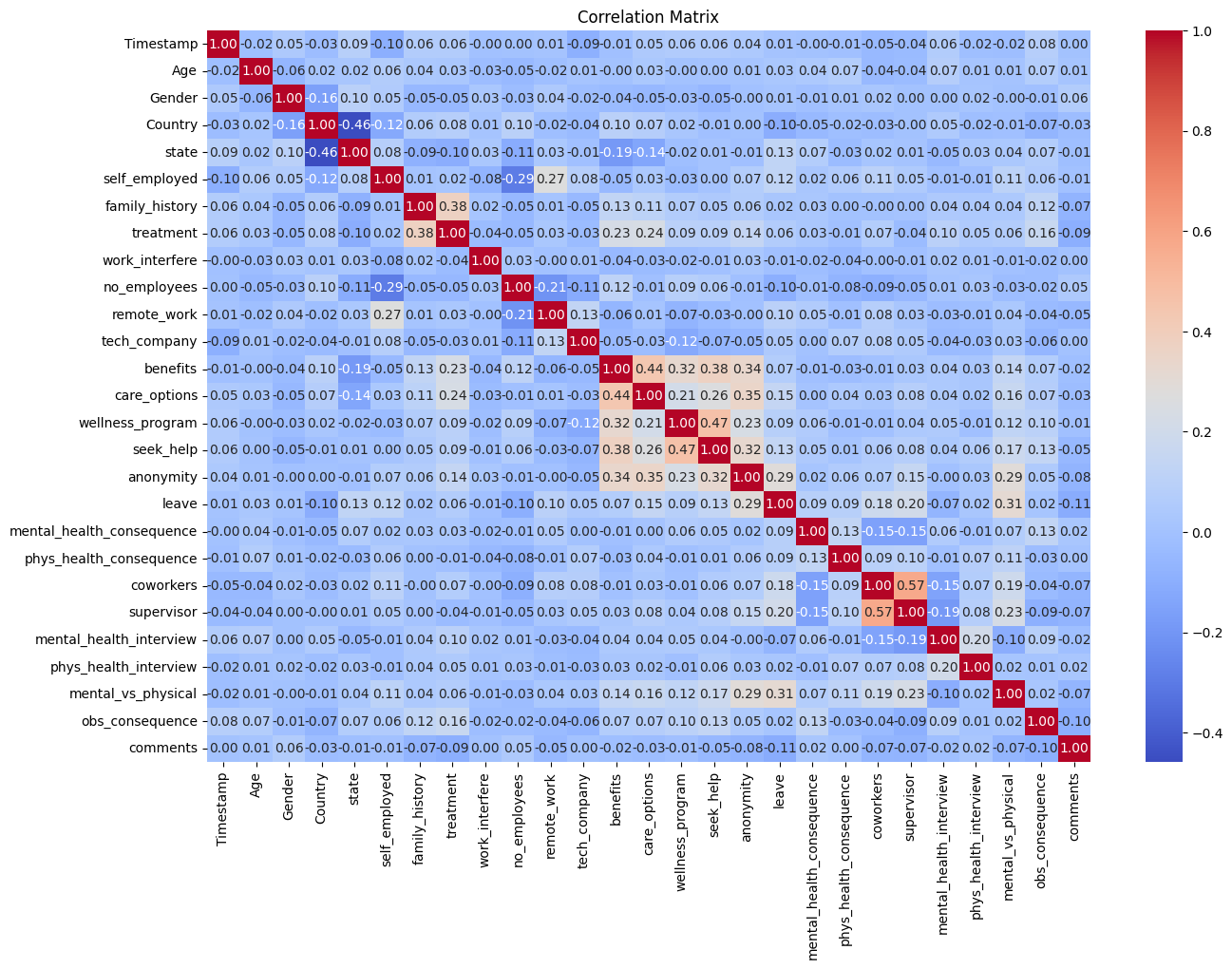
****

Figure : Correlation Analysis of features related to Personal data, Geography, Medical History, Care Options

### **Dimensionality Reduction Strategies**

The aim of dimensionality reduction was to enhance interpretability and reduce computational costs:

* **PCA (Principal Component Analysis)**:
  + Achieved approximately 95% variance capture with the first three components, thereby simplifying the feature space while preserving essential patterns.
  + Identified clusters where remote work and workplace benefits significantly impacted treatment-seeking behavior.

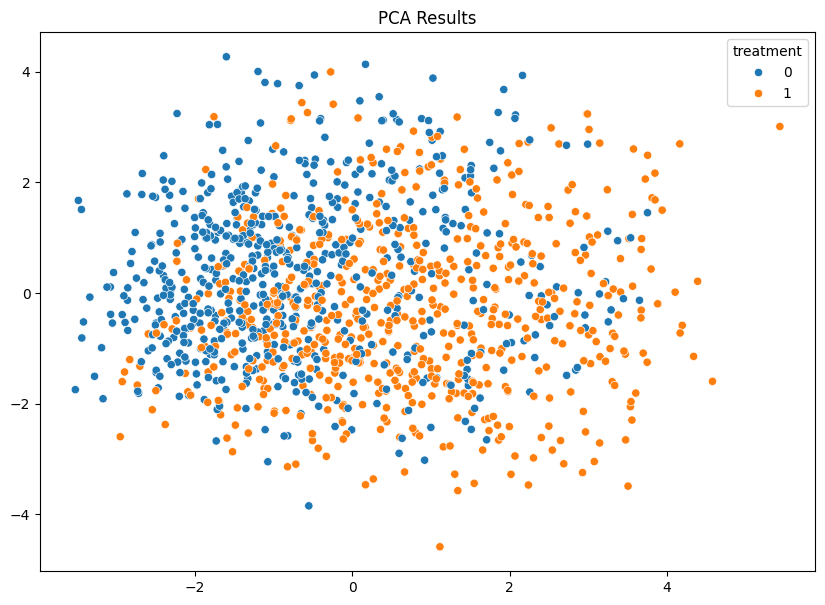


Figure : Principal Component Analysis (PCA) Results

* **UMAP and t-SNE**:
  + Proved effective in visualizing high-dimensional relationships, uncovering distinct clusters of individuals based on mental health outcomes.
  + UMAP demonstrated clearer separability between treatment-seeking and non-treatment-seeking groups, facilitating better cluster interpretation.

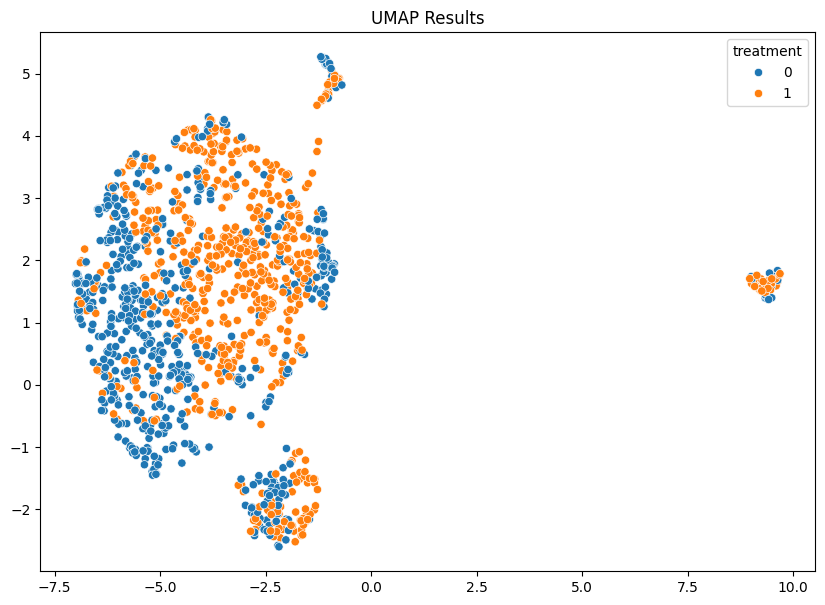


Figure : Uniform Manifold Approximately & Projection (UMAP) Results

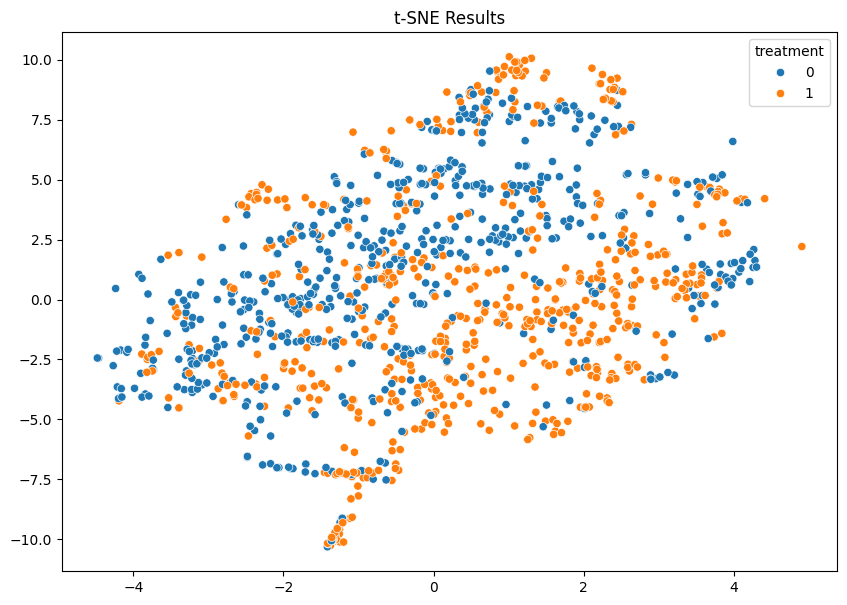


Figure : t-distributed Stochastic Neighbor Embedding (t-SNE) Results

### **Unsupervised Learning Techniques**

Clustering provided valuable exploratory insights into the dataset:

* **K-Means Clustering**:
  + The optimal cluster count was determined to be three, based on the elbow method.
  + ***Cluster Analysis:***
    - **Cluster 0**: Exhibited a high likelihood of seeking treatment, influenced by supportive workplace factors.
    - **Cluster 1**: Showed moderate treatment-seeking behavior, potentially highlighting issues such as stigma or resource gaps.
    - **Cluster 2**: Displayed minimal treatment-seeking despite experiencing high workplace stress, indicating critical areas for intervention.
* **Key Observations**:
  + The clustering analysis identified actionable segments for targeted employer strategies, especially for Cluster 2.

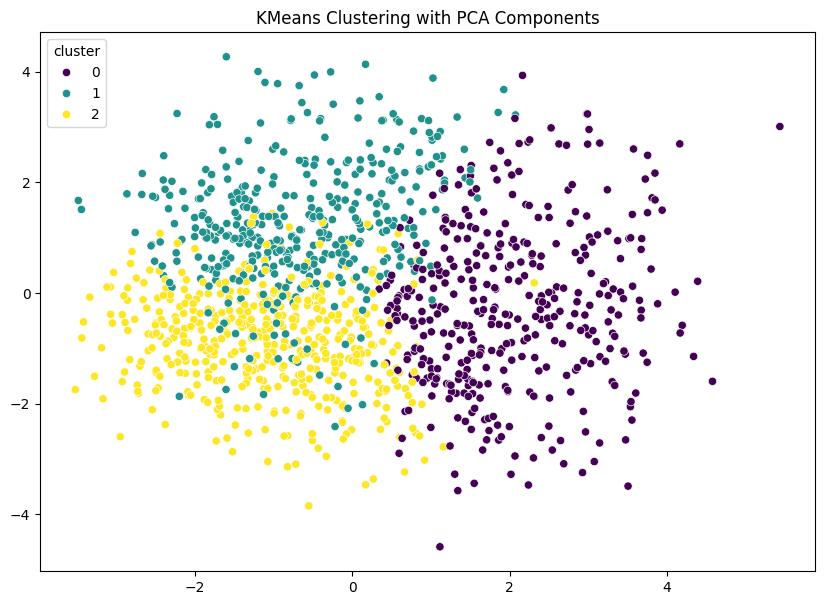


Figure : K-Means Clustering with PCA Components

### **Cross-Validation Strategy**

To ensure a robust and reliable performance evaluation, a stratified K-Fold cross-validation approach was implemented, focusing on key metrics and balancing class representation throughout the process.

* **Implementation**:
  + *Maintained class balance across all folds:* This approach was designed to prevent any over-representation of non-treatment-seeking responses, ensuring that each fold had a proportional representation of classes. This stratification was crucial in maintaining the integrity and validity of the evaluation.
  + *Evaluated models on key metrics:* Models were rigorously assessed using precision, recall, F1-score, and accuracy. By employing these metrics, the evaluation provided comprehensive insights into the model’s performance, highlighting both its strengths and areas for improvement.
* **Insights**:
  + *Consistent performance across folds:* The application of this approach consistently demonstrated the reliability of the model. The results across different folds remained stable, reinforcing the model's dependability.
  + *Variance in precision and recall:* This observed variance pointed out potential areas where the model might be overly reliant on specific features. This insight is critical for further refinement and enhancement of the model, ensuring its robustness and effectiveness across varied datasets.

### **Model Training & Performance Analysis**

* **Baseline Logistic Regression**:
  + The model achieved a balanced accuracy rate of 71.83%, demonstrating moderate levels of precision and recall.
  + Implemented regularization techniques, specifically L1 and L2 regularization, to minimize overfitting and ensure a stable performance across various datasets.
* **Random Forest Classifier**:
  + The Random Forest Classifier significantly outperformed the logistic regression model, achieving an impressive accuracy rate of 83.33%.
  + It was particularly effective in reducing false negatives, which is crucial for accurately identifying treatment-seeking individuals.
  + Metrics:
    - Class 0 (No Treatment):
      * Precision: 88%
      * Recall: 78%
      * F1-Score: 83%
    - Class 1 (Treatment):
      * Precision: 80%
      * Recall: 89%
      * F1-Score: 84%

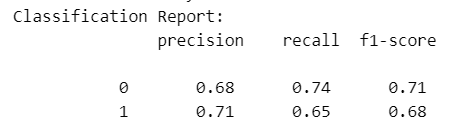


Table : Classification Report of Random Forest Classifier

### **Improvements & Results**

* **Hyperparameter Tuning**:
  + We meticulously adjusted the parameters of the Random Forest model, paying close attention to tree depth and the number of feature splits. This tuning process aimed to optimize the model's recall rate while ensuring that the precision was not compromised. By fine-tuning these hyperparameters, we enhanced the model's ability to identify treatment-seeking individuals effectively.
* **Feature Selection**:
  + To reduce the model's complexity and improve its interpretability, we concentrated on high-impact variables such as family history and supervisor support. This selective approach helped in maintaining a streamlined model that was easier to manage and interpret. As a result of this targeted feature selection, the validation accuracy post-reduction stood at 69.05%, which indicated a balanced trade-off between simplicity and performance.
* **Regularization**:
  + For the logistic regression model, we applied regularization techniques to stabilize the results. This process was crucial in maintaining balanced metrics across different classes, ensuring that the model did not overfit and provided consistent performance. The application of regularization contributed significantly to the robustness of the logistic regression model, enhancing its reliability across various datasets.

By focusing on these key improvements, the overall performance of our models was notably enhanced, demonstrating a thoughtful balance between complexity, precision, and stability.

### **Recommendations**

1. **Preferred Model**: The Random Forest model should be deployed as the primary model due to its demonstrated robustness and superior accuracy. Its ability to manage complex interactions and provide reliable predictions makes it an ideal choice for real-world implementation.
2. **Action Points for Employers**:
   * *Enhance Mental Health Resources and Supervisor Training:*
     + Employers should focus on improving mental health resources available to employees and providing comprehensive training to supervisors. It will foster a supportive workplace environment, crucial for employees' well-being and productivity.
   * *Address Barriers to Treatment-Seeking:* 
     + Employers need to address significant barriers such as stigma and lack of anonymity, which often prevent individuals from seeking necessary treatment. By tackling these issues, employers can create a more supportive and accepting work environment, encouraging more employees to seek the help they need.
3. **Future Work**:
   * *Incorporate External Datasets:*
     + It is essential to include external datasets as it will enhance the model’s applicability across different populations and settings.
   * *Explore Advanced Models:*
     + To achieve more precise & reliable predictions, future research should explore advanced models, such as boosting techniques, as it will further refine and improve prediction accuracy.
4. **Dashboard Integration**:
   * *Develop Real-Time Analytics Dashboard:*
     + The tool would visualize mental health trends and deliver actionable insights, facilitating targeted interventions. Serving as a powerful resource, the dashboard would enable continuous monitoring and enhancement of workplace mental health support.

**Dataset**: <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey>

**Project Link:**