Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Data-Driven Mental Health Analysis of Young Adults Using Clustering Algorithms

An Unsupervised Machine Learning Approach for Early Risk Identification

Janhavi Sachin Sudake
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce and Science College
Pimpri, Pune, Maharashtra, India

Shilpesh Sharad Dodake
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce and Science College
Pimpri, Pune, Maharashtra, India

***Abstract* - Academic pressure, emotional stress, and rapid lifestyle changes have led to a significant increase in mental health concerns among young adults. Traditional mental health assessment methods primarily rely on self-reporting and clinical evaluations, which often fail to uncover hidden behavioral patterns and early risk indicators. To address this limitation, this study presents a data-driven mental health analysis framework based on clustering algorithms. The proposed approach employs unsupervised machine learning techniques, specifically K-Means clustering, to group individuals based on psychological, behavioural, and academic attributes without relying on predefined labels. Data preprocessing and feature selection were performed to enhance clustering effectiveness, and the optimal number of clusters was determined using the Silhouette Score. The resulting clusters represent distinct mental health risk levels, enabling the early identification of vulnerable groups. This clustering-based framework supports proactive mental health intervention strategies and provides meaningful insights for counsellors and educational institutions. The system is scalable, ethically designed, and suitable for data-driven mental health support and preventive-care.**

Keywords* - *Mental Health Analysis, Clustering Algorithms, K-Means, Machine Learning, Student Well-being

I. INTRODUCTION

Mental health has emerged as a critical concern among young adults due to increasing academic demands, social pressures, and rapid lifestyle changes. Students often experience stress, anxiety, and emotional imbalance, which can significantly affect their academic performance, personal development, and overall well-being. Early identification of mental health risks is essential to prevent long-term psychological issues and promote timely intervention strategies.

Traditional mental health assessment methods largely rely on self-reporting, counselling sessions, and clinical evaluations. While these approaches are valuable, they are often limited by subjectivity, scalability challenges, and delayed detection of underlying mental health patterns. With the growing availability of digital mental health data, there is an increasing need for automated, data-driven approaches that can analyze large datasets and uncover hidden behavioral trends.

Recent advancements in machine learning have enabled the intelligent analysis of complex and high-dimensional data. In mental health research, most existing studies focus on supervised learning techniques for classification and prediction. However, such methods require labelled data, which are often difficult to obtain and may fail to capture emerging or previously unknown mental health patterns. This limitation highlights the importance of unsupervised learning methods, particularly clustering algorithms, in exploratory mental health analysis.

This study proposes a clustering-based framework for analyzing mental health data among young adults. By grouping individuals with similar psychological, behavioral, and academic characteristics, the proposed approach aims to identify distinct mental health risk profiles without relying on pre-defined labels. The insights generated through clustering can support early risk identification, proactive intervention strategies, and data-informed decision-making by counsellors and educational institutions.

II. METHODOLOGY

The proposed methodology follows a structured data-driven pipeline designed to analyze mental health patterns among young adults using unsupervised machine learning techniques. The overall workflow includes data collection, pre-processing, feature selection, clustering, and cluster evaluation. Each stage was designed to ensure meaningful pattern discovery and reliable clustering outcomes.

A. Dataset Description

The dataset used in this study consisted of mental health–related attributes collected from young adults, including psychological, behavioral, and academic factors. The data captured indicators such as stress levels, emotional well-being, academic pressure, and lifestyle-related variables. These attributes collectively provide a comprehensive representation of an individual's mental-health profile. To ensure ethical considerations, the dataset did not include personally identifiable information. The data were used strictly for analytical purposes, focusing on identifying patterns rather than diagnosing individuals.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

B. Data Preprocessing

Data preprocessing is a crucial step in improving the quality and reliability of clustering results. The raw dataset may contain missing values, inconsistent entries, and categorical attributes unsuitable for direct use in machine learning models. The preprocessing steps included handling missing values, encoding categorical variables into numerical form, and normalizing numerical features to ensure uniform scaling. These steps reduce bias and prevent certain features from dominating the clustering process due to scale differences.

C. Feature Selection

Feature selection was performed to retain the most relevant attributes that contributed to mental health pattern identification. Redundant and less informative features were removed to reduce dimensionality and improve clustering efficiency. The selected features represent key psychological, behavioral, and academic characteristics, ensuring that the clustering algorithm captures meaningful relationships among individuals. This step enhances the interpretability and computational efficiency.

D. Clustering Using K-Means Algorithm

K-Means clustering was employed as the primary unsupervised learning technique for grouping individuals based on similarities in mental health attributes. The algorithm partitions the dataset into k clusters by minimizing the intra-cluster variance and maximizing the inter-cluster separation. Multiple values of k are evaluated to identify the optimal number of clusters. Each data point was assigned to the nearest cluster centroid based on distance measures, resulting in distinct mental health groupings that represented different risk levels.

E. Cluster Evaluation Using Silhouette Score

To evaluate the quality of clustering and determine the optimal number of clusters, the Silhouette Score was used as the primary evaluation metric. The Silhouette Score measures the similarity of a data point to its own cluster compared to other clusters, providing an objective assessment of cluster separation. Higher Silhouette Score values indicate better-defined and well-separated clusters. By comparing the scores across different values of k, the most suitable clustering configuration was selected for further analysis and interpretation.

III. RESULTS AND DISCUSSION

This section presents the results of the clustering analysis and evaluates the effectiveness of the proposed approach in identifying meaningful mental health patterns in young adults. The clustering outcomes were analyzed using the Silhouette Score, and the resulting groupings were discussed to highlight their practical relevance.

A. Clustering Performance Evaluation

To determine the optimal number of clusters, K-Means clustering was applied with different values of k. The Silhouette Score was used as the evaluation metric to assess the cluster quality and separation. Table I presents a comparison of the clustering performance for different cluster counts.

TABLE I. Clustering Model Comparison Using Silhouette Score

Algorithm	Number of Clusters (k)	Silhouette Score
K-Means	2	0.354
K-Means	3	0.277
K-Means	4	0.212
K-Means	5	0.201

As shown in Table I, the highest Silhouette Score was achieved when the number of clusters was set to $k = 2$. This indicates that the data points are better separated and more compactly grouped in this cluster configuration than in higher values of k. As the number of clusters increased, the Silhouette Score decreased, suggesting reduced cluster cohesion and overlap among clusters.

B. Interpretation of Clusters

Based on the optimal clustering configuration, the dataset was segmented into distinct mental health groups. These clusters represent individuals with similar psychological, behavioral, and academic characteristics. This grouping enables the identification of varying mental health risk levels without relying on predefined labels. The clusters can be interpreted as representing relatively lower- and higher-risk mental health profiles. Such differentiation is valuable for understanding the underlying behavioral patterns and supports the early identification of vulnerable individuals. The clustering results demonstrate the effectiveness of unsupervised learning in uncovering hidden structures in mental health data.

C. Discussion

The experimental results highlight the suitability of clustering algorithms for exploratory analyses of mental health. Unlike traditional supervised approaches, the proposed clustering-based framework does not require labelled data and can identify emerging or previously unknown mental health patterns.

The use of the Silhouette Score provides an objective mechanism for validating cluster quality, ensuring reliable and interpretable results. The findings suggest that clustering can serve as a valuable decision-support tool for counsellors and educational institutions by enabling data-driven insights and proactive intervention strategies to be developed. Overall, the results confirm that the proposed methodology effectively addresses the challenges identified in the Introduction.

IV. CONCLUSION

This study presents a data-driven framework for analyzing mental health patterns among young adults using clustering algorithms. By leveraging unsupervised machine learning

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

techniques, this study demonstrated how meaningful mental health groupings can be identified without relying on predefined labels. The use of K-Means clustering, combined with Silhouette Score–based evaluation, enabled an objective assessment of cluster quality and supported the identification of distinct mental health risk profiles.

These findings highlight the potential of clustering-based approaches for early risk identification and exploratory mental health analysis. Such data-driven insights can assist counsellors and educational institutions in understanding the underlying behavioral patterns and designing proactive intervention strategies. The proposed methodology is scalable, ethically designed, and suitable for analyzing large mental health datasets.

Future work may extend this research by incorporating longitudinal data, additional clustering techniques or hybrid models that combine unsupervised and supervised learning. Integrating real-time data sources and domain expert feedback could further enhance the accuracy and applicability of this framework. Overall, this study contributes to the development of intelligent, preventive, and data-driven mental health-support systems.

REFERENCES

[1] World Health Organization (WHO), *Mental health of adolescents*, World Health Organization, Geneva, Switzerland, 2021.

[2] A. K. Jain, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[3] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed., Morgan Kaufmann, San Francisco, CA, USA, 2012.

[4] S. Dua and X. Du, *Data mining and machine learning in healthcare*, Academic Press, Cambridge, MA, USA, 2018.

[5] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed., Springer, New York, NY, USA, 2009.

[7] N. Shatte, D. Hutchinson, and S. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[8] M. S. Hossain et al., "Analyzing student mental health using machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp. 1–8, 2020.