Clustering-based Analysis of Risk Profiles for Depression using Quality of Life Data

Willian Jorge Sousa Furtado[1] [a], Pedro Almir Martins Oliveira[2] [b], Rossana Maria Castro Andrade[3] [c], Evilasio Costa Junior[3] [d], Wilson Castro[3] [e], Victória Tomé Oliveira[3] [f], and Pedro de Alcântara Santos Neto[4] [g]

[1] *Federal Institute of Education, Science and Technology of Maranhão (IFMA), Caxias, MA, Brazil*
[2] *Federal Institute of Education, Science and Technology of Maranhão (IFMA), Pedreiras, MA, Brazil*
[3] *Federal University of Ceará (UFC), Fortaleza, CE, Brazil*
[4] *Federal University of Piauí (UFPI), Teresina, PI, Brazil*

Keywords: Depression, Clustering, Quality of Life, Wearable Devices, Mental Health Risk, Unsupervised Learning.

Abstract: Depression is a condition that affects thousands of people worldwide, significantly compromising Quality of Life (QoL). Given the urgency of early diagnosis and the need for continuous monitoring, this study employs non-invasive data from wearable devices and unsupervised learning (clustering) to segment mental health risk profiles. Using the Healful dataset and its QoL score, the research compares the effectiveness of different algorithms. The strategic choice of the clustering technique, prioritizing business interpretability over purely statistical cohesion metrics, enabled a more pragmatic risk segmentation. The model, configured with $K = 4$, achieved a significant variation of 48.03 points in the QoL metric, clearly differentiating four distinct risk profiles: High Risk, Moderate-High, Moderate-Low, and Low Risk. This characterization provides a robust analytical framework for optimizing digital health interventions.

1 INTRODUCTION

Depression is characterized as a pathology that affects millions of people worldwide (Vitali et al., 2025). In this way, according to data from the Pan American Health Organization (Organização Pan-Americana da Saúde, 2025), more than 300 million individuals suffer from this disorder, significantly impacting quality of life (QoL) and making it one of the leading causes of global disability (Oliveira et al., 2025).

Moreover, the World Health Organization (World Health Organization, 2023) emphasizes that this disorder goes beyond normal mood variations, severely compromising emotional well-being and quality of life. Therefore, its impacts extend beyond individual suffering, also affecting professional performance, interpersonal relationships, and, in more severe cases, potentially leading to suicide.

In the medical domain, disease diagnosis, especially early diagnosis, is of utmost importance for both physicians and patients (Yang et al., 2024). Consequently, given the urgency of early diagnosis and the need for continuous monitoring, the advent of wearable devices offers a promising solution for passive and objective risk screening. According to (Sabry et al., 2022), wearable devices include any body-mounted equipment capable of capturing non-invasive physiological signals through various types of sensors. Furthermore, these devices enable the collection of large volumes of multimodal health and activity data (Yang et al., 2024), which can subsequently serve as proxies to identify changes in mental health states.

However, the challenge of monitoring through wearable devices lies in processing raw, often clinically unlabeled data. Unlabeled data are ubiquitous in many fields, including medicine, where large amounts of information are frequently collected without explicit labels. Therefore, unsupervised learning techniques are beneficial for analyzing medical data and discovering patterns and relationships that can assist in diagnosis, treatment, and drug discovery (Yang et al., 2024). In this context, unsupervised learning through clustering is a valuable technique for segmenting homogeneous

[a] https://orcid.org/0009-0006-2425-8550
[b] https://orcid.org/0000-0002-3067-3076
[c] https://orcid.org/0000-0002-0186-2994
[d] https://orcid.org/0000-0002-0281-2964
[e] https://orcid.org/0009-0007-4753-9926
[f] https://orcid.org/0000-0002-1400-522X
[g] https://orcid.org/0000-0002-1554-8445

risk groups, enabling the discovery of inherent risk profiles within the population.

In this context, the present study aims to segment mental health risk profiles, with a specific focus on identifying factors associated with depression. The segmentation is based on a subset of behavioral and socioeconomic features, previously validated through a comprehensive literature review on risk factors for the development of depression. Using the public Healful dataset, designed to assess Quality of Life (QoL) through data collected from wearable devices, along with its QoL metric (*psy ref score*), this research compares the effectiveness of different clustering algorithms, including K-Means, DBSCAN, Mean Shift, and Gaussian Mixture Models. To prioritize interpretability, the Fuzzy C-Means algorithm was selected. The study culminates in the characterization of four distinct risk profiles, providing an analytical map for optimizing future digital health interventions.

This is outlined as follows: Section 2 presents relevant background related to our study; Section 3 describes our material and methods; Sections 4 and 5 bring our results and discussion; and, finally, Section 6 concludes this paper.

2 Background

This section presents the theoretical foundations of data clustering analysis and the algorithms used in this study. Clustering aims to identify natural groupings within a set of patterns, points, or objects (Jain, 2010). The following subsections briefly describe the five clustering algorithms compared in this research.

2.1 K-Means

The K-Means algorithm takes as input the parameter k, corresponding to the desired number of clusters, and partitions a set of n objects into k groups, ensuring that the intra-cluster similarity is high and the inter-cluster similarity is low (Castro and Ferrari, 2016). K-Means starts with an initial partition into k clusters and assigns patterns to clusters to minimize squared error. Since the squared error always decreases with an increasing number of clusters k, it can only be minimized for a fixed number of clusters (Jain, 2010).

2.2 Fuzzy C-Means

The Fuzzy C-Means method is an extension of the K-Means algorithm in which each object has a degree of membership relative to each cluster. Unlike hard clustering algorithms, where an object either belongs or does not belong to a given group, in Fuzzy C-Means, an object may belong to multiple clusters simultaneously, but with varying degrees of membership (Castro and Ferrari, 2016).

2.3 Mean Shift

Mean Shift is a centroid-based algorithm that operates by iteratively updating candidate centroids so that they converge toward the mean of the points within a given region (Scikit-learn, 2024). The algorithm identifies clusters in datasets where the number of clusters is unknown. It finds clusters by iteratively shifting data points toward the densest regions of the feature space (Chugh, 2024).

2.4 DBSCAN

DBSCAN is a density-based algorithm designed to discover clusters and noise in datasets, formalizing the idea that the density of points inside a cluster is significantly higher than outside it (Ester et al., 1996). According to (Ester et al., 1996), the algorithm operates using two global parameters:

- **Eps:** The maximum radius that defines the neighborhood of a point. The ε-neighborhood of a point p includes all points q in the dataset whose distance $dist(p,q)$ is less than or equal to ε.
- **MinPts:** The minimum number of points that must exist within the ε-neighborhood of a point for it to be considered a *core point*.

2.5 Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a clustering technique used in unsupervised learning to determine the probability that a data point belongs to a given cluster (Carrasco, 2024). It assumes that a finite mixture of Gaussian distributions with unknown parameters generates all data points. Mixture models can be viewed as a generalization of K-Means clustering, incorporating information about the data covariance structure and the centers of the latent Gaussian components (Scikit-learn, 2024).

3 METHODOLOGY

The methodology of this study was structured to enable the segmentation of mental health risk profiles based on behavioral and sociodemographic data collected from wearable devices. This section sequentially describes the steps taken, from dataset selection and characterization to data preprocessing, scaling, determining the optimal number of clusters, and evaluating the applied algorithms. The goal is to ensure the reproducibility of the experiment and transparency in the justification of methodological choices.

3.1 Dataset Description

The present study used a public dataset available on the Kaggle platform (Almir, 2023).This dataset was originally compiled for the investigation of Quality of Life (QoL) and presents the following data collection structure:

Data Collection and Participants: The data were collected in two distinct phases. The first phase included 20 participants (from March to June 2022), and the second phase recruited 24 new participants between October 2022 and January 2023, composed primarily of undergraduate students.

Selection Criteria: Participants aged between 18 and 65 years were recruited by convenience sampling, with the essential requirement of being familiar with the use of wearable devices (smartwatches or fitness bands) and having the availability to wear them continuously.

Data and Variables: Daily IoHT (Internet of Health Things) data were collected anonymously and transmitted to the cloud. Every week, participants completed the WHOQOL-BREF questionnaire (covering physical and psychological domains), enabling the construction of the Quality of Life metric (psychological reference score).

Final Result: After preprocessing, the dataset was consolidated into 1,373 instances.

Ethics: Ethics: The research process was approved by the Ethics Committee of the Federal University of Ceará (UFC) on March 9, 2022, under protocol number 56153322.0.0000.5054 (legal opinion number 5.282.056).

The use of this dataset is crucial for the study's objectives, as it provides the necessary combination of rich behavioral features and a validated Quality of Life metric. This integration allowed the analysis of QoL to be extended toward the inference of mental health risk profiles, associating the decline in QoL with the potential risk of developing depressive symptoms.

3.2 Data Analysis

Remote monitoring of mental health is highly important for individuals with mental disorders, as well as for their caregivers and clinicians (Sheikh et al., 2021). In light of this, the clustering methodology, an unsupervised learning approach, was adopted in this study due to its intrinsic suitability for mental health risk screening and dynamic monitoring.

The choice of unsupervised learning is strategic, as it enables the discovery of inherent risk profiles based solely on the manifestation of behavioral patterns observed in the dataset's features. It is important to emphasize that this technique is particularly advantageous in healthcare settings where formal clinical labels are limited, inconsistent, or time-consuming to obtain (Yang et al., 2024). Therefore, this approach is essential for establishing transparent and interpretable risk levels that can be used to prioritize clinical or preventive interventions.

3.3 Feature Selection

The initial analytical strategy involved cross-validating public datasets to correlate depression risk with the features available in the Healful public Quality of Life (QoL) dataset. To this end, an extensive search was conducted across multiple public data repositories, including Kaggle[1], Mendeley Data[2], Data.World[3], and Zenodo[4]. The search utilized keywords such as *"depression"*, *"mental health"*, and *"wearable data"* to identify potential datasets of interest.

The purpose of this approach was to identify a dataset that simultaneously met two essential criteria:

1. Contained a reliable and well-documented depression diagnosis label; and
2. Included a set of features comparable in granularity to the behavioral (wearable) and sociodemographic metrics present in the dataset.

This cross-validation step was expected to provide an external benchmark for evaluating the clustering approach's generalizability and robustness.

3.3.1 Search for Datasets

However, the conducted search revealed no datasets that satisfied both requirements simultaneously. The inability to find a suitable complementary dataset for cross-validation was primarily attributed to inconsistencies and limitations in the available data. These limitations can be categorized into three main issues:

Variable Incompleteness: Several datasets included some relevant features, but were insufficient for clustering. For instance, some contained only a single socioeconomic feature (*e.g.*, gender or income), which restricted their analytical potential.

[1] https://www.kaggle.com

[2] https://data.mendeley.com

[3] https://data.world

[4] https://zenodo.org

Lack of Reference Label: Other datasets offered behavioral data of interest but lacked clinically validated depression labels, which are necessary for segmentation validation.

Inconsistency in Clinical Labeling: In some instances where depression labels were available, their documentation was ambiguous or nonexistent. Moreover, some datasets used arbitrary scoring systems (e.g., 5, 20, or 48) without formal descriptions or clinical justification for the scoring methodology.

In conclusion, none of the identified datasets met the requirements for conducting a reliable cross-dataset analysis. Consequently, an alternative validation approach was required to ensure the methodological soundness of the study.

3.3.2 Methodological Validation of Features

Given the lack of success in identifying a public dataset suitable for cross-validation, a methodological transition was undertaken, in which the validity of the features was established through secondary studies on risk factors for the development of depression.

The literature review encompassed studies that analyzed populations of young and adult individuals across different demographic contexts, identifying multiple relevant risk factors. The studies considered included (Sousa, 2022), (Ndikumana et al., 2025), (Lisznyai et al., 2014), (Torres et al., 2013), (Purborini et al., 2021), (Rahman and Kohli, 2024), (Kader Maideen et al., 2014), (Al Balawi et al., 2019) and (Mokona et al., 2020).

The purpose of this process was to ensure that the final clustering subset of the Healful dataset comprised only variables whose influence on depression risk was scientifically validated. Table 1 summarizes the main risk factors identified and their corresponding selected features.

Table 1: Validated Risk Factors in the Literature and Corresponding Selected Features

Validated Risk Factors in the Literature	**Selected Features from the Healful Dataset**
Gender; Marital status (divorced/single); Financial and employment situation; Income; Physical activity; Social isolation; Low self-esteem; Stress	steps; gender; income; maritalstatus_single; maritalstatus_married; differentwifi; differentlocations; profession_fulltimeworker; profession_parttimeworker; profession_selfemployed

Ultimately, this process resulted in a final subset of ten features that composed the clustering scope. This careful selection ensured that the segmentation of mental health risk profiles was grounded in solid empirical evidence, thereby enhancing the interpretability and scientific rigor of the analytical results.

3.4 Data Preprocessing and Scaling

Data preprocessing is a fundamental step in preparing the dataset for clustering algorithms. Datasets often contain several issues, such as missing values, noise (attributes with incorrect values), features with low predictive value, and class imbalance (Batista, 2003). The process was divided into two main phases: *(i)* data cleaning and standardization, and *(ii)* scaling and outlier handling, as described below.

3.4.1 Data Preprocessing

To ensure data quality and consistency, a rigorous data cleaning and preprocessing procedure was performed on the dataset. This process was applied systematically within the clustering model to prepare the dataset for analysis and machine learning model training, ensuring the data were in an appropriate, consistent format.

The initial stage focused on standardizing data types and identifying missing values. Columns were converted to numeric data types, with non-numeric values replaced by *Not a Number (NaN)*, a fundamental step in ensuring successful data cleaning before analyses or model training. The treatment of missing values and data preparation were then customized to meet the specific requirements of the different clustering algorithms.

3.4.2 Data Scaling and Outlier Treatment

Data scaling was essential to ensure that distance calculations, upon which all algorithms in this study are based, were not dominated by variables with large magnitudes, such as *steps* and *income*.

Initially, the *Standard Scaler* was used. However, due to the observed asymmetry and outliers in the behavioral and financial data, this method proved suboptimal, yielding low Silhouette scores. The Standard Scaler's sensitivity to the mean and standard deviation distorted the data distribution, masking the actual cluster structure.

To mitigate this issue, the method was replaced with the *Robust Scaler*, which subtracts the median and divides by the Interquartile Range (IQR) (Castro and Ferrari, 2016). This methodological adjustment proved crucial, significantly increasing cluster separation and cohesion, thereby validating the Robust

Scaler's superiority for this dataset.

3.5 Determining the Number of Clusters (K)

In this section, the procedure for determining the optimal number of clusters (K) in the clustering models is described. Section 3.3.1 presents the method for estimating the initial value of K, while Section 3.3.2 discusses the final selection process and the rationale for the configuration adopted in this study.

3.5.1 Elbow Method

The *Elbow Method* was employed to estimate the optimal number of clusters (K) for algorithms requiring prior initialization of this parameter, such as *K-Means* and *Fuzzy C-Means*.

The procedure was executed using data scaled by the *Robust Scaler*, ensuring that all features contributed equally to the computation of the Euclidean distance. Figure 1 illustrates the relationship between the number of clusters and the Within-Cluster Sum of Squares (WCSS).

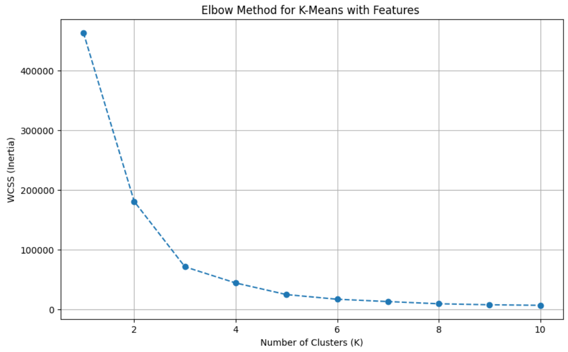

Figure 1: Relationship between the number of clusters and the WCSS.

The analysis of the WCSS curve indicates that the most pronounced inflection point (the "elbow") occurs at $K = 3$. From this point onward, the marginal reduction in WCSS becomes less significant, suggesting that $K = 3$ represents the most efficient structure for segmenting the risk profiles.

3.5.2 Final Selection of K

Although the statistical rigor of the *Elbow Method* initially indicated $K = 3$ as the optimal number of clusters, suggesting a triadic segmentation into low, medium, and high risk profiles, the final decision to adopt $K = 4$ was made in order to increase the granularity and level of detail in profile characterization.

The validity of this expanded structure was supported by the analysis of Quality of Life (QoL) score distributions, which showed that the transition from three to four clusters enabled the subdivision of the original highest risk cluster into two distinct subgroups. This adjustment was essential to isolate a specific High Risk profile, characterized by severely low QoL scores, thereby allowing a more precise identification of groups with more urgent clinical needs.

As a result, the four cluster configuration enabled the construction of a more actionable risk scale for digital health interventions, categorized into *Low Risk*, *Moderate Low Risk*, *Moderate High Risk*, and *High Risk*.

3.6 Model Evaluation

Five clustering algorithms were tested: K-Means, Fuzzy C-Means, Mean Shift, DBSCAN, and Gaussian Mixture Model (GMM), whose theoretical foundations are described in Section 2. Each algorithm was evaluated based on its ability to segment risk profiles according to the Quality of Life score for psychological domain, using the Silhouette Coefficient as the primary performance indicator.

3.6.1 Silhouette Coefficient and Selection Criterion

The performance and separation quality of the clustering algorithms were evaluated using the *Silhouette Coefficient*. This metric can be interpreted as a measure of how well an object fits within its assigned cluster, reflecting the cohesion and separation between clusters (Castro and Ferrari, 2016).

The index ranges from $[-1.0, +1.0]$, where values close to $+1.0$ indicate that the objects are well clustered (high cohesion) and well separated from other clusters.

For the *DBSCAN* and *Mean Shift* models, whose clusters are generated automatically, the parameters were calibrated to replicate the structure of $K = 4$ clusters defined by the business criterion. For *DBSCAN*, the score was calculated only for the data points that were not classified as noise (label -1).

3.7 Model Selection

The selection of the final algorithm was based on a comparative analysis between mathematical cohesion and the capacity to discriminate risk profiles. Although the Silhouette coefficient is the standard metric for evaluating clustering, the results demonstrated that a high Silhouette score does not always translate into useful segmentation for the mental health domain.

The Mean Shift algorithm achieved the highest Silhouette Coefficient recorded in the study (0.8206), which theoretically would indicate an ideal configuration. However, when analyzing the mean Quality of Life (QoL) metric (*psy_ref_score*) for each generated cluster, a critical failure in profile separation was observed (Table 2).

Table 2: QoL Mean Distribution - Mean Shift ($K = 4$)

Cluster	Mean psy_ref_score
0	61.63
1	66.75
2	66.00
3	66.00

As demonstrated, Mean Shift concentrated most of the variation within an extremely narrow range of only 5.12 points. This overlap renders the segmentation ineffective, as clusters 1, 2, and 3 are practically indistinguishable from a clinical perspective, failing to identify increased risk groups. In contrast, the Fuzzy C-Means (FCM) algorithm, despite presenting a lower Silhouette Coefficient (0.3874), demonstrated a superior capacity to discriminate extreme profiles.

Table 3: QoL Mean Distribution - Fuzzy C-Means ($K = 4$)

Cluster	Mean psy_ref_score
0	35.21
1	60.21
2	66.66
3	83.24

The FCM configuration achieved a separation amplitude of 48.03 points on the Quality of Life scale. Unlike Mean Shift, FCM accurately isolated Cluster 1 (35.21), representing the critical vulnerability group. This refined segmentation capability is essential for the clear identification of extreme risk groups, which justifies the selection of FCM over Mean Shift based on its greater practical utility and clinical relevance to the proposed model's objectives.

4 RESULTS

This section presents the results from applying the clustering algorithms described in the methodology. Initially, the overall performance of the models is compared using the Silhouette Coefficient (Section 4.1). Subsequently, Section 4.2 provides a detailed characterization of the risk profiles identified in the study.

4.1 Overall Comparison

The performance results are based on the *Silhouette Coefficient*, which evaluates both intra-cluster cohesion and inter-cluster separation. Accordingly, Table 4 summarizes the comparative performance of the algorithms under two analysis configurations: $K = 3$ (suggested by the Elbow Method) and $K = 4$ (adopted to enhance business interpretability and granularity). For consistency, all results presented were obtained using the *Robust Scaler*, which proved to be the most stable preprocessing technique.

Table 4: Comparative Performance of Clustering Algorithms (Using Robust Scaler).

Algorithm	Silhouette Score (K=4)
Mean Shift	0.8206
Fuzzy C-Means	0.3941
K-Means	0.4976
DBSCAN	0.7877
GMM	-0.0540

As shown in Table 4, the *Mean Shift* algorithm achieved the highest Silhouette score, followed closely by *DBSCAN*. These two algorithms, therefore, exhibited the strongest cluster cohesion and separation among all the tested models. In contrast, *K-Means* and *Fuzzy C-Means* achieved moderate performance, while the *Gaussian Mixture Model (GMM)* displayed poor clustering behavior with a negative score. Consequently, the subsequent analysis focused primarily on the algorithms that demonstrated both numerical stability and interpretability.

4.2 Risk Profile Characterization (Fuzzy C-Means, K = 4)

Once the optimal algorithm and cluster configuration were determined, the next step was to characterize the resulting four risk profiles. This characterization was based on the mean values of the QoL score for psychological domain and the behavioral and sociodemographic variables associated with each cluster. Consequently, this section presents both the average QoL per cluster and a descriptive summary of the defining attributes of each risk profile.

Table 5 displays the average QoL values and their respective risk labels, revealing a total variation of 48.03 points between the lowest and highest scores. This substantial range confirms that the clusters are well differentiated in terms of perceived quality of life, reinforcing the validity of the adopted segmentation.

Table 5: Average QoL score for psychological domain and Risk Profile Classification (Fuzzy C-Means, K = 4).

Risk Label	Mean QoL score
High Risk	35.21
Moderate-High Risk	60.21
Moderate-Low Risk	66.67
Low Risk	83.24

As observed, the distribution of mean QoL values follows a clear ascending order across the four groups, which supports a coherent and interpretable progression from high to low risk.

5 Discussion

This study prioritized the interpretability of the model outputs. The Mean Shift algorithm achieved the best statistical performance, with the highest Silhouette coefficient (0.8200). However, the analysis of the Quality of Life score for psychological domain showed that the model was ineffective at differentiating risk profiles, as the variation between clusters was small (approximately 5 points).

Given this scenario, the Fuzzy C-Means (FCM) algorithm was selected as the final model. Although it presented a lower Silhouette coefficient (0.3874), FCM provided a substantially more informative separation, with an amplitude of approximately 50 points in the QoL metric. This segmentation proved more suitable for analyzing depression risk, allowing the identification of patterns consistent with the mental health literature.

The following paragraphs present the interpretative discussion of the four identified risk profiles:

Profile 1: High Risk (Cluster 0). This group represents the highest vulnerability to depression, characterized by the combination of low physical activity and high routine instability. These patterns indicate behavioral disorganization and possible signs of isolation or lack of daily regularity, both widely associated with an increased likelihood of developing depressive symptoms (Torres et al., 2013).

Profile 2: Moderate-High Risk (Cluster 1). Although composed of physically active individuals, this profile shows elevated risk due to psychosocial factors. The predominance of single individuals and the high load of full-time work may reflect loneliness and emotional overload, two factors frequently reported in the literature as triggers for depressive symptoms (Torres et al., 2013; Al Balawi et al., 2019).

Profile 3: Moderate-Low Risk (Cluster 2). Defined by financial instability and high mobility, this group reflects precarious living conditions and economic insecurity. These elements frequently contribute to chronic stress and increased risk of depression (Al Balawi et al., 2019; Rahman and Kohli, 2024; Lisznyai et al., 2014).

Profile 4: Low Risk (Cluster 3). This group represents the reference profile of well-being and protection against depression. It is characterized by high financial and professional stability and a higher proportion of married individuals, suggesting greater emotional and social support, which are recognized protective factors against depressive disorders (Sousa, 2022; Rahman and Kohli, 2024; Al Balawi et al., 2019).

Overall, the obtained segmentation demonstrates that depression risk can be indirectly inferred through behavioral patterns and socioeconomic variables, reinforcing the potential of unsupervised learning in mental health risk screening.

6 FINAL REMARKS

This study successfully achieved its objective of segmenting mental health risk by analyzing user profiles from the Healful Quality of Life (QoL) dataset, using a subset of features validated by nine studies on depression risk factors. The results demonstrate that data-driven unsupervised methods can meaningfully capture behavioral and sociodemographic patterns associated with mental health vulnerability.

From a methodological perspective, the clustering-based approach proved effective for analyzing the Quality of Life score for psychological domain. The strategic decision to prioritize business interpretability over purely statistical optimization guided the selection of the *Fuzzy C-Means (FCM)* algorithm with the configuration $K = 4$. Although this model had a relatively lower Silhouette Coefficient (0.3941), it showed the most significant variation in the QoL metric (48.03 points). This degree of differentiation is fundamental for distinguishing risk profiles and for translating findings into actionable insights for mental health monitoring.

The main contribution of this research lies in the detailed characterization of four distinct risk profiles, namely *High Risk*, *Moderate-High Risk*, *Moderate-Low Risk*, and *Low Risk*. Specific behavioral and demographic drivers shape each profile. For instance, the *High-Risk Profile* was associated with a paradoxical pattern of high total sleep time yet low overall QoL, emphasizing the importance of focusing on *sleep quality rather than duration*. Conversely, the *Low-Risk Profile* displayed *low sleep efficiency* as its primary vulnerability factor, suggesting that even among healthier individuals, subtle behavioral indicators may reveal latent risks.

In conclusion, the segmentation obtained in this study validates the proposed methodology for identifying population subgroups at elevated risk of developing depression. The four established profiles provide a robust analytical framework for guiding preventive and therapeutic interventions with greater precision. Furthermore, the ability to discriminate between groups with clear clinical relevance, supported by the wide variation in the QoL score for psychological domain, ensures that healthcare resources can be directed more efficiently toward individuals most in need. Consequently, this approach holds significant potential for enhancing the personalization and impact of future digital mental health strategies aimed at improving overall Quality of Life.

REFERENCES

Al Balawi, M. M., Faraj, F., Al Anazi, B. D., and Al Balawi, D. M. (2019). Prevalence of depression and its associated risk factors among young adult patients attending the primary health centers in tabuk, saudi arabia. *Open Access Macedonian Journal of Medical Sciences*, 7(17):2908.

Almir, P. (2023). Self-reported quality of life dataset. Accessed: 2025-01-09.

Batista, G. E. A. P. A. (2003). *Pré-processamento de dados em aprendizado de máquina supervisionado*. Dissertação (mestrado em ciência da computação), Universidade de São Paulo, São Paulo.

Carrasco, O. C. (2024). Gaussian mixture model explained. Accessed: 2025-11-09.

Castro, L. N. d. and Ferrari, D. G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*. Saraiva, São Paulo.

Chugh, V. (2024). Mean shift clustering: A comprehensive guide. Accessed: 2025-11-09.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Kader Maideen, S. F., Mohd. Sidik, S., Rampal, L., and Mukhtar, F. (2014). Prevalence, associated factors and predictors of depression among adults in the community of selangor, malaysia. *PLOS ONE*, 9(4):e95395.

Lisznyai, S., Vida, K., Németh, M., and Benczúr, Z. (2014). Risk factors for depression in the emerging adulthood. *The European Journal of Counselling Psychology*, 3(1):54–68.

Mokona, H., Yohannes, K., and Ayano, G. (2020). Youth unemployment and mental health: prevalence and associated factors of depression among unemployed young adults in gedeo zone, southern ethiopia. *International Journal of Mental Health Systems*, 14(1):61.

Ndikumana, F., Izabayo, J., Kalisa, J., Nemerimana, M., Nyabyenda, E. C., Muzungu, S. H., and Sezibera, V. (2025). Machine learning-based predictive modelling of mental health in rwandan youth. *Scientific Reports*, 15(1):16032.

Oliveira, P. A. M., Andrade, R. M., Santos Neto, P. A., Santos, I. S., Junior, E. C., and Oliveira, V. T. (2025). Internet of health things and machine learning for continuous quality of life monitoring. *Health and Quality of Life Outcomes*, 23(1):92.

Organização Pan-Americana da Saúde (2025). Depressão. Accessed: 2025-11-01.

Purborini, N., Lee, M. B., Devi, H. M., and Chang, H. J. (2021). Associated factors of depression among young adults in indonesia: A population-based longitudinal study. *Journal of the Formosan Medical Association*, 120(7):1434–1443.

Rahman, M. A. and Kohli, T. (2024). Mental health analysis of international students using machine learning techniques. *PLOS ONE*, 19(6):e0304132.

Sabry, F., Ahmed, S., Mostafa, A., and Mohamed, E. (2022). Machine learning for healthcare wearable devices: The big picture. *Journal of Healthcare Engineering*, pages 1–25.

Scikit-learn (2024). Clustering. Accessed: 2025-11-09.

Sheikh, M., Qassem, M., and Kyriacou, P. A. (2021). Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in Digital Health*, 3:1–13.

Sousa, C. D. (2022). *Análise dos fatores de risco associados à depressão no Brasil, no ano de 2019*. Monografia (bacharelado em estatística), Universidade Federal Fluminense, Niterói.

Torres, E., March, S., Socias, I. M., and Esteva, M. (2013). Factores de riesgo de síndrome depresivo en adultos jóvenes. *Actas Españolas de Psiquiatría*, 41(2):84–96.

Vitali, E., Cattane, N., D'Aprile, I., Petrillo, G., and Cattaneo, A. (2025). Systemic inflammation at the crossroad of major depressive disorder and comorbidities: A narrative review. *International Journal of Molecular Sciences*, 26(19):9382.

World Health Organization (2023). Depressive disorder (depression). Accessed: 2025-11-02.

Yang, L., Amin, O., and Shihada, B. (2024). Intelligent wearable systems: Opportunities and challenges in health and sports. *ACM Computing Surveys*, 56(7):1–42.