Predictive Models in Mental Health Based on Unsupervised Data Clustering

Inoc Rubio Paucar[1], Cesar Yactayo-Arias[2], Laberiano Andrade-Arenas[3]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Departamento de Estudios Generales, Universidad Continental, Lima, Perú[2]
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú[3]

Abstract—**In the university context, students' mental health has been progressively affected over time. The objective of this research was to develop a predictive model of machine learning based on the K-Means algorithm, with the purpose of identifying and classifying mental health profiles among university students. For the construction of this model, the standard Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was applied, which encompasses five stages: business understanding, data understanding, data preparation, modeling, and evaluation. The results obtained suggest that the generated clusters produce consistent groupings in key variables such as screen time, hours of sleep, and level of physical activity, allowing the characterization of different student profiles. This approach provides valuable information for designing academic support strategies and programs aimed at students' well-being and mental health. The early identification of behavioral patterns and lifestyle habits enables educational institutions to implement preventive and personalized measures, fostering improved academic performance and university adaptation.**

Keywords*—*Behavioral patterns; clustering; machine learning; mental health; university students

I. Introduction

In the field of higher education, mental health has emerged as a critical issue in recent years due to the notable increase in disorders such as anxiety, depression, and chronic stress. University students constitute a particularly vulnerable population, as they face multiple challenges inherent to this formative stage, including academic pressure, adaptation to new environments, uncertainty regarding professional futures, financial difficulties, and, frequently, separation from family support networks. According to the World Health Organization (WHO), approximately 20% of young people experience some type of mental disorder, many of which begin precisely during university life [1] [2]. In this respect, it becomes evident that this problem demands attention in every social context, as it exacerbates personal difficulties among university students worldwide.

Mental health in the university population represents both a scientific and medical issue of growing relevance [3]. From a medical perspective, mental disorders such as anxiety, depression, and chronic stress not only compromise students' emotional well-being but also directly affect their physical health, quality of life, and academic performance. If not detected or treated in time, these disorders may lead to serious consequences such as university dropout, deterioration of social relationships, and, in the most severe cases, suicide attempts. From a scientific perspective, this issue poses significant challenges for interdisciplinary research [4] [2]. Experts in medicine consider mental health problems as a set of conditions that comprehensively affect student life, also impacting the family environment.

There is an urgent need to better understand the individual, social, and contextual risk factors that contribute to the onset of these disorders during university years. Despite the rising prevalence, many cases remain underdiagnosed due to stigmatization, lack of institutional resources, or limited awareness of the importance of preventive psychological care. Furthermore, the scientific study of student mental health requires the development and validation of effective tools for early diagnosis, symptom monitoring, and implementation of personalized interventions. The integration of approaches based on emerging technologies, such as artificial intelligence and machine learning, opens new possibilities for risk detection and prediction; however, gaps in the evidence remain and must be addressed through rigorous research [5] [6]. Technologies thus represent a viable option within the broader social challenge, constituting an impactful factor for prevention depending on the context under study.

This research is justified by the need to reduce the incidence of psychological disorders and other problems that significantly affect university students, negatively impacting the achievement of their academic goals and their professional future. In this regard, it is essential to develop effective mechanisms to counteract these difficulties, thereby promoting students' overall well-being and supporting their persistence and success in higher education. Faced with this issue, it becomes crucial to implement efficient mechanisms that enable the timely identification of at-risk students in order to design and implement personalized intervention strategies [7] [8]. In this context, the use of machine learning techniques represents an innovative and powerful tool, as it allows the analysis of large volumes of data and the discovery of hidden patterns that would not be detectable through traditional methods. Specifically, this research proposes the use of clustering algorithms, such as K-means, to group students according to relevant psychological, academic, and social characteristics. This methodology will allow the segmentation of the student population into groups with similar risk levels, facilitating informed decision-making by educational institutions [9]. In this way, more effective support strategies can be developed, aimed at improving mental health, student retention, and academic success, thus contributing to students' overall well-being and to the strengthening of the higher education system.

The objective of this research is to develop a predictive model based on clustering techniques and machine learning, aimed at identifying and classifying mental health profiles among university students through the analysis of multivariate data related to lifestyle habits, academic characteristics, and emotional factors. The purpose of this model is to facilitate the early detection of potential psychological disorders and to contribute to the design of preventive strategies or personalized interventions that enhance student well-being and academic performance.

II. Literature Review

This section presents the structure of the literature review, which is divided into two main parts. The first addresses the theoretical foundations, providing a detailed discussion of the mental health variable and the clustering approach, with emphasis on the K-means algorithm. The second part comprises the related works, analyzing previous studies conducted by various authors who have applied machine learning techniques, specifically clustering algorithms, to the study and analysis of mental health among university students. The purpose of this review is to provide the theoretical basis for the research and to highlight the relevance and applicability of the proposed approach.

A. Theoretical Bases

1) Mental health: Mental health in university students refers to the emotional, psychological, and social state that directly influences their overall well-being and academic performance. This stage represents a critical period in personal development, where students face multiple challenges such as academic pressure, adaptation to new environments, family expectations, and, in many cases, financial limitations. In particular, some students must finance their studies without family support, which significantly increases their emotional and financial burden [10] [11]. This situation of solitude and economic responsibility can generate high levels of stress and anxiety, affecting their mental health and hindering the achievement of their academic goals. Table I presents the key factors influencing the mental health of university students, detailing the causes and the associated emotional and psychological effects, such as stress, anxiety, and feelings of isolation, which may impact their well-being and academic performance.

2) Clustering: Clustering is an unsupervised data analysis technique aimed at identifying and grouping objects or instances into subsets known as clusters. Its fundamental principle is that elements within the same group share a high degree of internal similarity, while exhibiting marked differences from elements in other groups. The evaluation of this similarity or distance is carried out using various metrics, including Euclidean distance, Manhattan distance, and correlation coefficients, selected according to the characteristics and nature of the data. Unlike supervised methods, clustering does not require predefined labels, which makes it a powerful exploratory tool for uncovering hidden patterns, latent structures, and intrinsic relationships in large volumes of information [12] [13]. Its versatility has positioned it as an essential resource in areas such as market segmentation, bioinformatics, image analysis, and anomaly detection. Table II shows the representation of three metrics for evaluating clusters: Euclidean Distance measures the closeness between points; the Silhouette Score (0.65) indicates good group separation; and the Calinski–Harabasz Index (200) reflects well-defined and separated clusters.

B. Related Work

Several studies have indicated that multiple factors related to university students' mental health can lead to different illnesses. The authors [14] [15] highlight the growing concern regarding mental health problems in university students. In particular, Singh evaluated various machine learning techniques, such as classification methods, clustering, and a hybrid neural network, applied to an academic dataset from the engineering faculty. The results showed that the hybrid algorithm, by combining classification and clustering approaches, significantly outperformed the other models, achieving high accuracy in predicting both academic performance and students' mental well-being. Unlike the previous study, this author addressed the influence of mental, emotional, social, and spiritual health on the academic performance of university students. To this end, data were collected through a questionnaire administered to 214 undergraduate students, and machine learning techniques such as feature selection, regression, neural networks, Naïve Bayes, and multidimensional analysis were employed, along with tools such as SPSS, R, and Python. The analysis revealed that these methods were effective in predicting academic performance and identifying the psychosocial factors with the greatest impact during the transition to virtual and hybrid learning modalities. On the other hand, the author [16] proposed an intervention program aimed at improving the mental health of university students, using a Bayesian model for the dynamic updating of parameters. Cluster random sampling was also used to divide participants into an experimental group and a control group. The results showed that the intervention produced positive changes in key aspects such as mental health literacy, self-efficacy in coping with psychological problems, and behavioral attitudes toward seeking help. In another study, an author developed a strategy to classify stress in engineering students using wearable devices that recorded signals during the Montreal Imaging Stress Task (MIST). Using machine learning models, the study managed to classify stress as resting, moderate, and high, achieving accuracies of 99.98% with EEG signals and 99.51% with Electrodermal Activity (EDA), Heart Rate (HR), and Skin Temperature (SKT), outperforming previous studies and demonstrating the effectiveness of these devices for real-time monitoring [17].

In a relevant study on university students' mental health, anxiety levels and their causes were investigated using a Likert-scale-based questionnaire administered to 127 participants. Based on the collected data, machine learning models such as Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM) were trained. The results were promising, with accuracies of 71.05% for Naïve Bayes and Decision Tree, 78.9% for Random Forest, and 75.5% for SVM, showing the potential of these algorithms to effectively predict anxiety levels and provide valuable insights into their causes. In another investigation, the application of predictive models for addressing mental health issues

TABLE I. FACTORS AFFECTING THE MENTAL HEALTH OF UNIVERSITY STUDENTS

Factor	Description	Impact on mental health
Academic pressure	Demands and high expectations in school performance	Stress, anxiety, exhaustion
Adaptation to new environments	Changes in social life and university environment	Loneliness, uncertainty, anxiety
Family expectations	Pressure to meet family goals	Anxiety, fear of failure
Economic limitations	Difficulty in affording studies without family support	Financial stress, constant worry
Individual responsibility	Emotional and financial burden of handling everything alone	Increased stress, feeling of isolation

TABLE II. COMMON METRICS TO EVALUATE CLUSTERING QUALITY

Metric	Description	Example value
Euclidean Distance	Measures the direct distance between points	-
Silhouette Score	Quality of clustering (from -1 to 1)	0.65
Calinski-Harabasz Index	Separation between clusters and internal cohesion	200

in university students employed approaches such as neural networks and support vector machines. These models were key in documenting a bibliometric analysis, whose results revealed that Artificial Intelligence and Machine Learning show a consistent growth trend in addressing mental health problems among young university students [18] [19]. The lack of timely attention to mental health problems increases the risk of worsening over time. This study collected data through questionnaires on depressive, anxiety, and obsessive-compulsive symptoms, applying models such as multiple linear regression to identify associations and the random forest algorithm for predictions. The results indicated that 20% of students showed severe depression and/or suicidal ideation, that financial concerns were strongly associated with depression, and that random forest showed high accuracy in predicting the maintenance of well-being (balanced accuracy = 0.85) and absence of suicidal ideation, but low accuracy in detecting worsening cases (balanced accuracy = 0.49). Furthermore, in another study on mental health in university students, it was identified that cognitive and somatic symptoms of depression were the most influential variables, with a high negative predictive value (0.89) and practically null positive predictive value, highlighting stress as a key parameter for mental well-being. Through surveys administered to 144 students from various universities, multilayer perceptrons (MLP) combined with Principal Component Analysis (PCA) and Grid Search Cross-Validation (GSCV) were used, achieving an accuracy of 80.5%, precision of 1.000, F1-score of 0.890, and recall of 0.826. The findings revealed that 11.26% presented high social stress and 24.10% extremely high psychological stress, although the study was limited by the use of non-probabilistic sampling and self-reported data [20] [21].

On the other hand, stress is a key factor in understanding the prevalence of mental health problems, which is why this study applied automated machine learning methods to identify its main causes. A survey was conducted with 355 students from 28 universities, using the Boruta algorithm to select the most relevant predictive factors. The models evaluated included decision tree, random forest, and support vector machine, and their performance was measured using confusion matrix, ROC curves, and k-fold cross-validation. The results highlighted that heart rate, systolic and diastolic blood pressure, sleep quality, smoking habits, and academic history were the most influential variables in predicting stress, with the random forest model achieving the best performance, reaching an accuracy of 89.72% and an area under the curve (AUC) of 0.8715, outperforming the other techniques in accuracy and predictive capacity. An important factor is social media, which is a determinant of mental health; thus, the author proposed evaluating its influence through the collection and analysis of 66,000 student posts. The research focused on the use of natural language processing models to identify symptomatic expressions of depression, anxiety, and stress, also employing the SARIMA algorithm to forecast the number of mental health consultations on campus. The incorporation of social media data significantly improved the accuracy of these predictions, achieving a correlation of r = 0.86 and a symmetric mean absolute percentage error (SMAPE) of 13.30, representing a 41% improvement compared to models that did not consider social information, thereby demonstrating the key value of social media in detecting and monitoring student mental well-being [22] [23]. Another study conducted with university students using multicenter cross-sectional surveys aimed to compare the performance of different machine learning models, including k-nearest neighbors, naïve Bayes, neural networks, and random forest. The results showed that algorithms based on Random Forest achieved the highest accuracy in identifying negative indicators of mental well-being. Among the most relevant variables for prediction were the number of weekly sports activities, body mass index, academic average, sedentary hours, and age. These findings provide valuable recommendations for modernizing and optimizing mental health assessment and monitoring, both at the individual level and within the university environment. In relation to the previous study, the research developed the Emotion Based Mental Health Classifier (EMHC), a system that combines sentiment analysis and categorical data to classify students' mental health into three levels: good, moderate, and poor. Based on data from a web application applied in universities in the NCR region of India, it uses advanced algorithms such as DeepFace, K-means, and Support Vector Classifier. This innovative tool facilitates the assessment and monitoring of student mental well-being in a global context of increasing psychological burden [24] [25].

In contrast to other studies, a multidimensional feedback approach was proposed based on a clustering analysis applied to data from 174 university students. Using the k-means algorithm, participants without severe symptoms were segmented into three differentiated groups, while those with severe symptoms received professional care recommendations. This personalized feedback not only improved mental health literacy but also fostered a sense of belonging, reducing resistance to seeking support and enabling timely intervention [26]. Another important factor regarding alcohol consumption and perceived stress in Ecuadorian university students was analyzed using 7,134 records obtained from the Alcohol Use Disorders Identification Test (AUDIT-C) questionnaire, the PSS-10 scale, and sociodemographic data. By applying Lanz's (2013) methodology and using k-means and hierarchical clustering algorithms, three distinct groups were identified that reflect combined patterns of alcohol consumption and stress levels, thus supporting the theory linking both factors and providing a complementary perspective to traditional statistical analysis. Finally, another study sought to raise awareness of the mental health status of students at Satya Wacana Christian University, using the K-Means machine learning algorithm to identify emotional patterns based on numerical responses from 32 students. The analysis, conducted with Orange3 and based on the silhouette index, formed three groups: one associated with depression, mainly including students from the 2018 cohort and some from 2020; another group reflecting mental prosperity, composed of students from the 2018, 2019, and 2020 cohorts; and a third group in a state of harmony, with students from various cohorts between 2017 and 2019. The results show that the year of admission influences mental well-being, with greater prosperity observed in earlier cohorts, higher depression among third-year students, and a group that achieves emotional balance across different cohorts [27] [28]. In summary, related work presents important limitations that open up opportunities for further research: most are based on unrepresentative samples, which restricts the generalizability of the findings; these could be improved with more representative data and the best use of more advanced techniques.

III. Methodology

A. Definition of the CRISP-DM Methodology

CRISP-DM methodology is a widely used standard for guiding data mining and machine learning projects. Its purpose is to provide a structured, flexible, and non-linear framework that enables the development of models and the extraction of useful knowledge from data. This approach is characterized by being iterative and adaptable, meaning that its phases can be repeated or feed back into one another. In this way, the process can be continuously redesigned to respond to changing business needs, ensuring that solutions generate tangible value aligned with the strategic objectives of the organization [29] [30]. Additionally, Fig. 1 illustrates the CRISP-DM methodology for data analysis, which begins with information collection and preparation, followed by problem definition and data understanding using descriptive statistics. Variables are subsequently transformed for cluster modeling (K-Means), evaluating metrics such as silhouette and inertia, and finally, the resulting clusters are interpreted and validated. The process defined by this methodology represents a flexible form of implementation in machine learning projects, allowing its adaptation according to the specific topic and objectives of each case. Fig. 2 illustrates the architecture of the methodology and internally describes each process, showing how the model is developed according to the technical specifications.

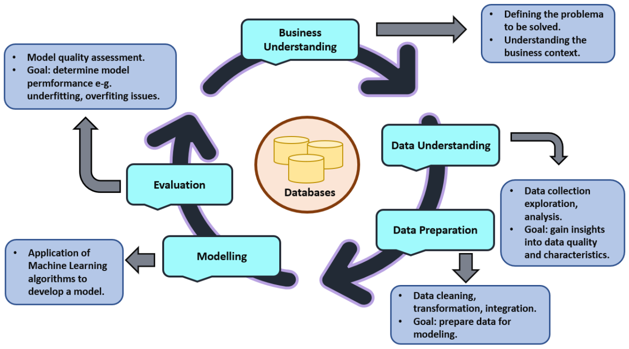

Business Understanding
Defining the problema to be solved.
Understanding the business context.
Model quality assessment.
Goal: determine model permformance e-g. underfitting, overfiting issues.
Data Understanding
Databases
Evaluation
Data collection exploration, analysis.
Goal: gain insights into data quality and characteristics.
Data Preparation
Modelling
Application of Machine Learning algorithms to develop a model.
Data cleaning, transformation, integration.
Goal: prepare data for modeling.

Fig. 1. CRISP-DM methodology.

1) Business understanding: In this section, corresponding to the first stage of the CRISP-DM methodology, the specific criteria necessary to address the problem are analyzed. Based on the data obtained, the problem to be solved is posed, which involved abstracting the real situation and transforming it into a clear and achievable analysis [31]. According to Table III, information is presented from a database downloaded from Kaggle, organized into records containing variables related to individuals' personal characteristics and habits. Fields include data such as gender, age, weight, height, type of transportation used, consumption of fast food and alcoholic beverages, among other factors. This information is valuable because it allows for the analysis of behavioral patterns linked to nutrition, physical activity, and lifestyles, which can be used in public health studies, obesity prediction, or in the development of preventive strategies to improve the population's quality of life.

2) Data understanding: For this section, certain data analysis concepts were applied using the Python programming language, which allowed us to examine and explore the important variables for the proposed model to understand their importance, detect possible inconsistencies and analyze the relationship between variables, serving as a fundamental basis for preparing the data that will feed the algorithmic model [32]. In this regard, Table IV shows a statistical summary of the key variables analyzed in the sample of 1,000 students. Demographic data, lifestyle habits, and perceived stress levels are included. The average age of participants is 20.34 years, with values ranging from 15 to 26. On average, students spend 6.91 hours in front of screens per day, sleep 6.45 hours, and engage in 5.02 hours of physical activity per week. The stress level, coded from 1 (low) to 3 (high), averages 1.85, suggesting that most students fall between low and medium levels of stress. This summary provides an understanding of the distribution and variability of the variables studied, serving as a reference for subsequent analysis and the development of predictive models.

On the other hand, Fig. 3 of boxplots presents a summary view of the distribution of four key variables: screen time, sleep duration, physical activity, and stress level. Each box shows the interquartile range, with the median dividing the distribution in half. The whiskers extend to non-outliers, and individual

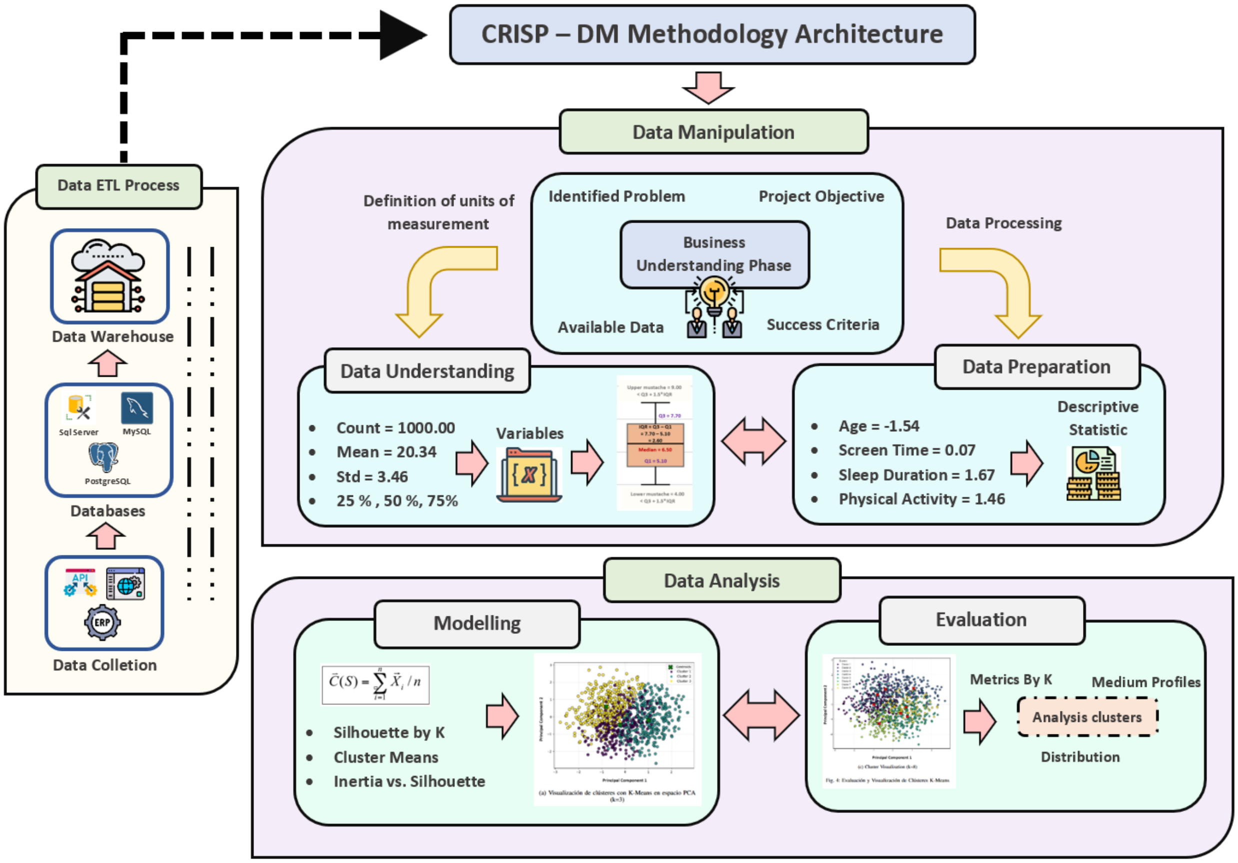


Fig. 2. Architecture CRISP-DM methodology.

TABLE III. BUSINESS UNDERSTANDING PHASE

Element	Description
Identified Problem	There are increasing mental health issues among university and high school students, reflected in factors such as stress, irregular sleep, excessive screen time, and low physical activity. This affects academic performance and overall well-being.
Project Objective	Use the information from the dataset (lifestyle habits, academic and emotional variables) to develop a predictive model based on clustering and machine learning techniques, which allows identifying students at risk or with similar mental health patterns.
Available Data	Demographics: name, gender, age, educational level.
	Lifestyle habits: screen time, sleep duration, weekly physical activity.
	Emotional factors: stress level, anxiety before exams.
	Observed outcome: change in academic performance.
Success Criteria	**Technical:** achieve well-defined and stable clusters (metrics such as Silhouette Score or Davies-Bouldin Index).
	Practical: obtain interpretable profiles that allow designing early intervention strategies or academic and emotional support.

TABLE IV. STATISTICAL SUMMARY OF THE STUDY VARIABLES

Stadistic	Age	Screen Time (hrs/día)	Sleep Duration (hrs)	Physical Activity (hrs/Week)	Stress Level
count	1000.00	1000.00	1000.00	1000.00	1000.00
mean	20.34	6.91	6.45	5.02	1.85
std	3.46	2.91	1.47	2.93	0.70
min	15.00	2.00	4.00	0.00	1.00
25%	17.00	4.40	5.10	2.60	1.00
50%	20.00	6.90	6.50	5.00	2.00
75%	23.00	9.50	7.70	7.60	2.00
max	26.00	12.00	9.00	10.00	3.00

points beyond them represent outliers. This graph reveals that all four variables exhibit data dispersion, with the presence of outliers that may indicate extreme behaviors or measurements in the studied sample, which warrants further investigation to understand why these data deviate from the main behavior.

3) Data preparation: According to this section, in the Data Preparation phase, raw data is transformed into high-quality data ready for modeling. This stage, which goes hand in hand with the Data Understanding phase, involves cleaning the data (such as handling missing and outlier values) and transforming categorical variables into numerical variables so that the model's algorithms can properly process them [33].

Table V presents a sample of student data with variables normalized for analysis. Each row represents a student, and the columns display their data relative to the average. Age, Screen Time, Sleep Duration, and Physical Activity are standardized values, where a positive value indicates the student has an above-average age, screen time, sleep duration, or physical activity, and a negative value indicates a lower one. Stress Level is categorized qualitatively (Low, Medium, High) and also represented numerically (Stress Level Num) for use in machine learning models. For example, the first student has a below-average age (-1.54), above-average sleep duration (1.67), and a medium stress level. Table VI summarizes the descriptive statistics for the variables in a dataset after normalization. Numerical variables such as Age, Screen Time, Sleep Duration, and Physical Activity now have a mean of 0 and a standard deviation of 1, indicating they have been standardized. Meanwhile, the categorical variable Stress Level was converted to the numerical variable Stress Level Num, with a mean of 2.16, showing that the data tend toward the "Medium" stress level.

4) Modelling: In this section, the proposed algorithmic model is developed to address the identified problem. Supported by the analysis carried out in the previous phases, statistical graphs, tables, and mathematical expressions are defined to justify the selection and application of the K-Means algorithm as a clustering technique [34].

Fig. 4a shows the two-dimensional projection of the students through a Principal Component Analysis (PCA), where the three clusters identified by the K-Means algorithm are represented. Each color corresponds to a different group, and the centroids are indicated with 'X' markers. This visualization allows for the appreciation of the relative separation between groups and the distribution of the data in the reduced space. On the other hand, Fig. 4b presents the variation of the Silhouette index for different values of K (number of clusters). The maximum point is reached at k = best k, indicating the best partition in terms of internal cohesion and separation between groups. Additionally, the mathematical formula of the Silhouette coefficient is included, which reinforces the interpretation of the criterion used to validate the optimal number of clusters.

As shown in Table VII, the clustering analysis provides a clear overview of the data distribution and group characteristics. Table (a) reports the Silhouette Scores for different values of k, highlighting that the clustering quality is highest at $k = 3$ with a score of 0.51. Table (b) presents the mean values of Screen Time, Sleep Duration, and Physical Activity within each cluster, allowing us to profile the groups: Cluster 0 exhibits moderate screen time and sleep, Cluster 1 shows lower screen time and higher physical activity, and Cluster 2 has higher screen time with lower physical activity. Finally, Table (c) summarizes the inertia values along with the Silhouette Scores for each k, providing insight into both the compactness and separation of the clusters. Overall, these tables collectively illustrate the optimal number of clusters and the characteristics of each cluster group.

B. K-means Mathematical Formulas

1) Distance to centroid: The distance between a point x_p and a cluster centroid μ_i is computed using Euclidean distance (see eq.1) [35].
This distance is used to assign points to the nearest cluster.

$$d(x_p, \mu_i) = \sqrt{\sum_{j=1}^{m} (x_{pj} - \mu_{ij})^2} \tag{1}$$

where, m is the number of dimensions. This distance is used to assign points to the nearest cluster and to compute intra-cluster cohesion.

2) Cluster assignment: Each point is assigned to the cluster with the nearest centroid (see Eq. 2).
Points are grouped with the closest centroid to ensure cluster cohesion [36].

$$C_i = \{x_p : \|x_p - \mu_i\|^2 \leq \|x_p - \mu_j\|^2, \forall j = 1, \ldots, k\} \tag{2}$$

where, C_i is the set of points in cluster i. This ensures that points are grouped with the closest centroid.

3) Centroid update: Centroids are recalculated as the mean of all points in the cluster (see Eq. 3).
This moves the centroid to the center of its cluster for better stability.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{3}$$

This adjustment moves the centroid to the center of its cluster, improving cluster cohesion.

4) Cost function (Inertia): K-means minimizes the sum of squared distances from points to their centroids (see Eq. 4).
Lower inertia indicates tighter clusters and better cluster compactness.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4}$$

Lower inertia means points are closer to their centroid, indicating tighter clusters.

5) Convergence criterion: The algorithm stops when centroids change minimally between iterations (see Eq. 5).
A small tolerance ϵ ensures that clusters are stable.

$$\|\mu_i^{(t+1)} - \mu_i^{(t)}\| < \epsilon, \quad \forall i \tag{5}$$

where ϵ is a small tolerance value. This ensures clusters are stable.

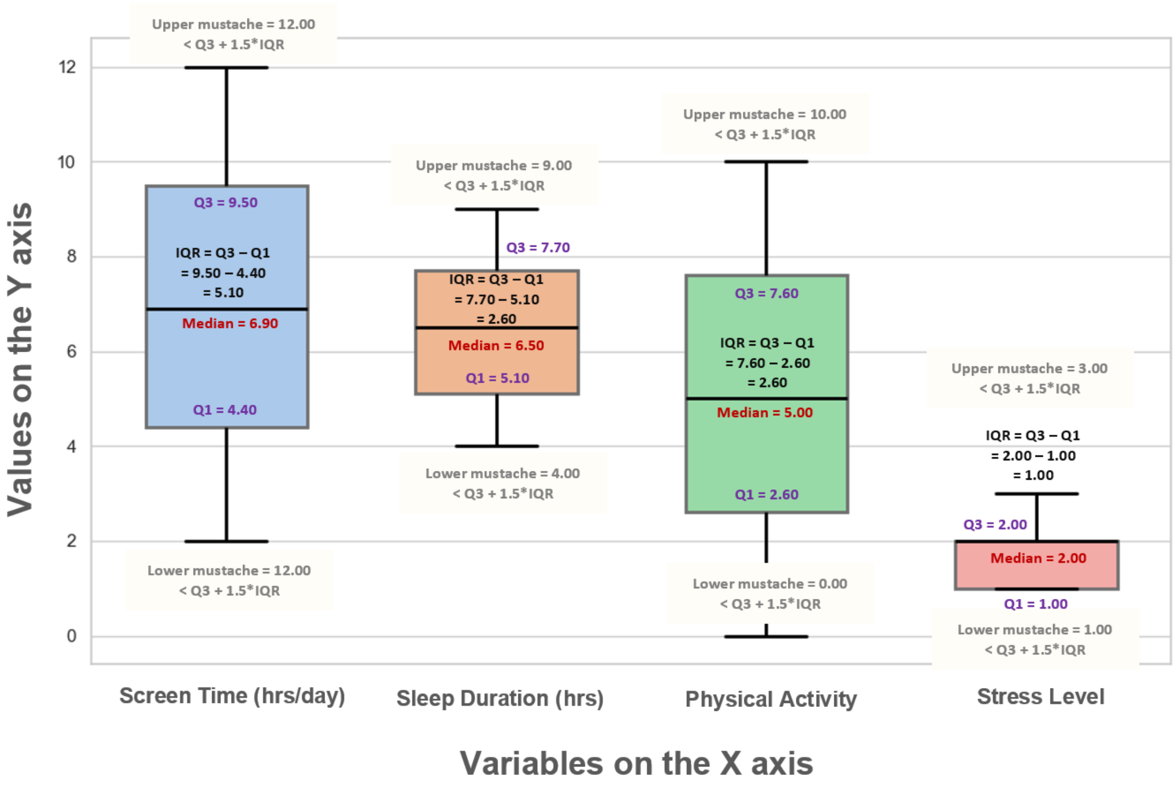


Fig. 3. Distribution of variables and the presence of outliers.

TABLE V. STUDENT DATA SAMPLE

Age	Screen Time (hrs)	Sleep Duration (hrs/week)	Physical Activity	Stress Level	Stress Level Num	Academic Performance	Anxiety Before Exams
-1.54	0.07	1.67	1.46	Medium	2.00	3.5	B
1.35	-1.24	-0.99	-1.64	Medium	2.00	2.8	C
-0.10	0.89	-0.72	0.40	Medium	2.00	3.2	B
-0.10	1.34	-0.58	0.16	High	3.00	2.1	D
-0.97	-1.41	-0.72	-0.65	Medium	2.00	3.9	A
0.50	0.20	-0.10	0.50	Low	1.00	4.1	A

TABLE VI. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

Variable	Mean	Standard Deviation	Minimum	Maximum
Age (Normalized)	0.00	1.00	-1.54	1.35
Screen Time (Normalized)	0.00	1.00	-1.41	1.34
Sleep Duration (Normalized)	0.00	1.00	-0.99	1.67
Physical Activity (Normalized)	0.00	1.00	-1.64	1.46
Stress Level Num	2.16	0.72	1.00	3.00

TABLE VII. (A) SILHOUETTE SCORES BY k, (B) CLUSTER MEANS, AND (C) INERTIA VS. SILHOUETTE

(a) Silhouette Scores by k

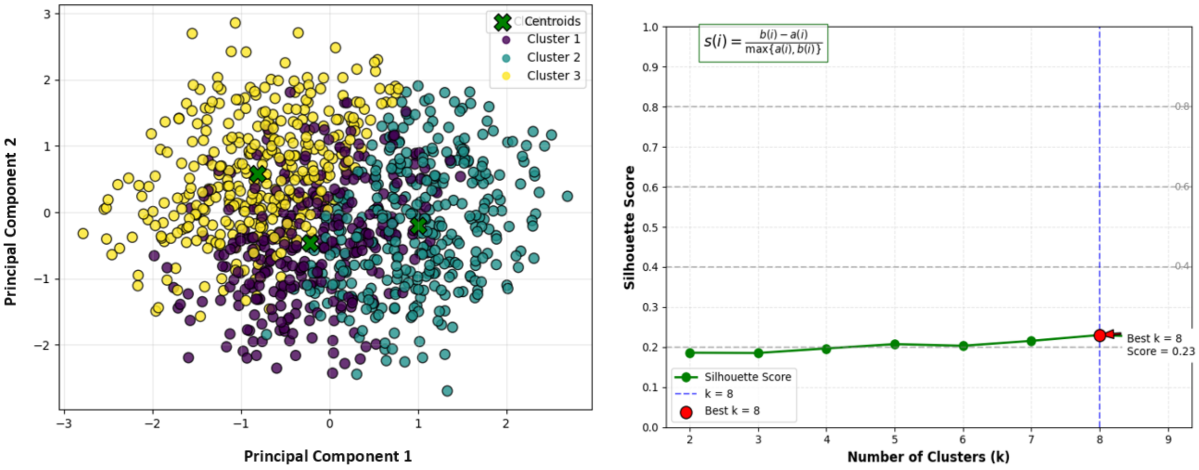
k	Silhouette
2	0.42
3	0.51
4	0.47
5	0.45
6	0.43

(b) Cluster Means (Example)

Cl.	Screen	Sleep	Activity
0	5.3	6.8	2.1
1	3.1	7.5	3.8
2	6.4	5.2	1.7

(c) Inertia vs. Silhouette

k	Inertia	Silhouette
2	580.1	0.42
3	430.4	0.51
4	390.2	0.47
5	350.1	0.45
6	320.5	0.43



(a) Visualizing clusters with K-Means in PCA space (k=3).

(b) Evaluating the optimal number of clusters with silhouette score.

Fig. 4. Student clustering using K-Means: PCA visualization and analysis of the optimal number of groups.

6) Total inertia (Optional): The total inertia quantifies overall cluster compactness (see Eq. 6).

It measures how well points fit within their clusters; lower values indicate tighter clustering.

$$J_{total} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{6}$$

where, k is the number of clusters, C_i the points in cluster i, x a data vector, and μ_i the cluster centroid.

7) Cluster prediction: Once the algorithm converges, a new point x_{new} is assigned to the cluster with the nearest centroid.

This completes the clustering process, assigning each point to its optimal cluster (see Eq. 7).

$$\hat{C} = \arg\min_{i} \|x_{\text{new}} - \mu_i\| \tag{7}$$

This completes the K-means clustering process, assigning each point to its optimal cluster.

IV. Result

In the Evaluation phase of the CRISP-DM process, the quality of the obtained model is verified, comparing its results with the business objectives and criteria defined in the initial phases. This process not only involves analyzing statistical and performance metrics, but also evaluating the consistency of the clusters generated by the K-Means algorithm with respect to the problem context. In this way, it is determined whether the model is suitable for practical implementation or if it requires additional adjustments to the parameters, input data, or the methodology used [37] [38].

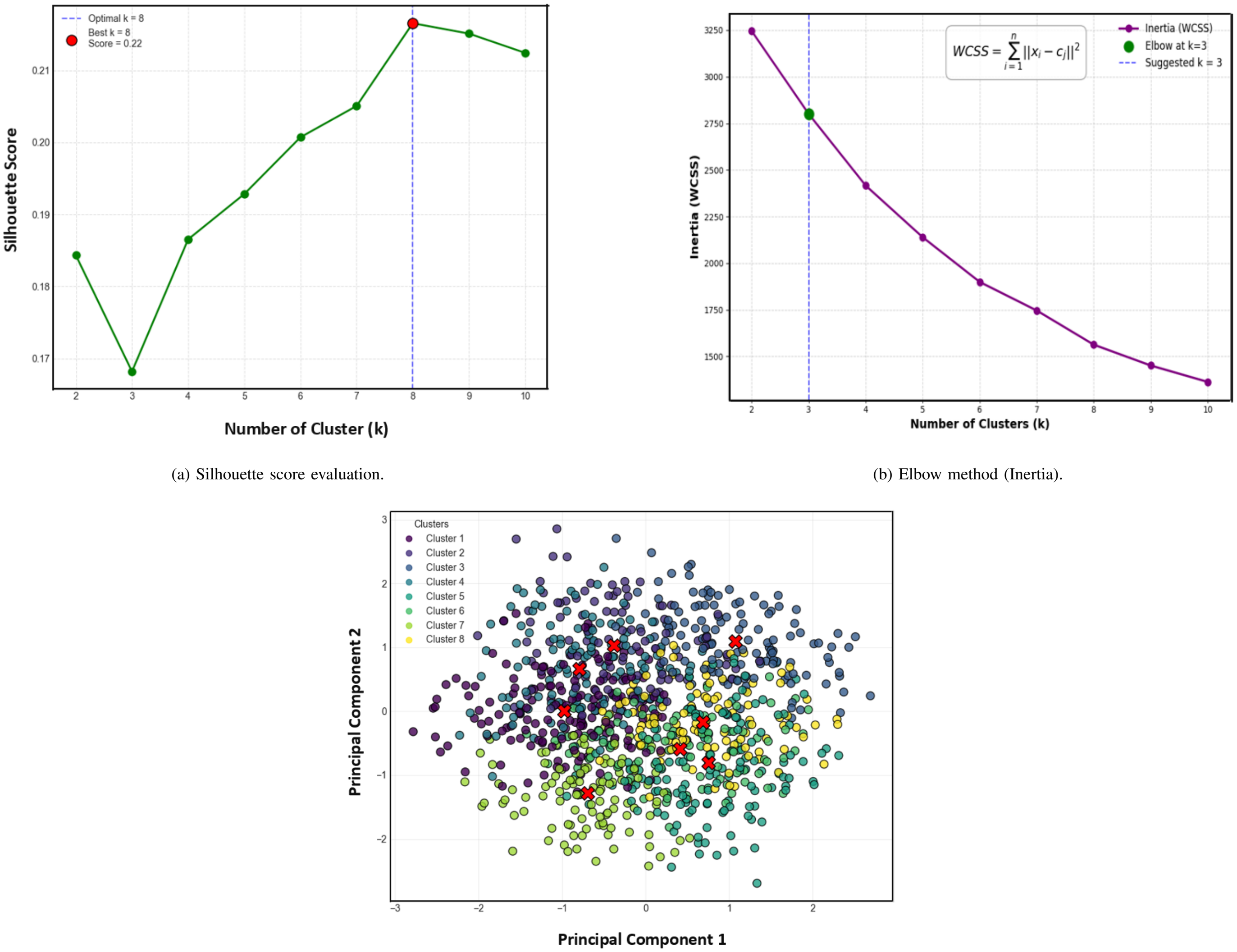
A. Evaluation of Result

Fig. 5 presents a comprehensive evaluation of the K-Means clustering analysis. Fig. 5a illustrates the Inertia curve (elbow method), which is used to identify the optimal number of clusters (k) by minimizing the sum of squared distances from the points to their centroids. Fig. 5b complements this evaluation with the Silhouette Score, a metric that measures cluster cohesion, where higher values indicate that objects are well grouped. Finally, Fig. 5c provides a visualization of the resulting clusters in a PCA (Principal Component Analysis) space, allowing observation of both the data distribution and the location of centroids for each selected cluster.

The references in the group of Tables VIII provide a comprehensive overview of the K-Means cluster analysis. Table VIIIa of Metrics by k shows that the Silhouette Score reaches its highest point at $k = 8$, suggesting that this is the optimal number of clusters for segmenting the data. Complementing this finding, Table VIIIb of Cluster Distribution demonstrates that the resulting 8 clusters have a relatively balanced size, with no groups being excessively large or small. Finally, the third and most descriptive, Table VIIIc of Cluster Profiles (Means), provides a detailed characterization of each group. For example, Cluster 1 is distinguished by having the highest average screen time but the lowest physical activity, while other clusters are defined by different combinations of age, sleep, and activity, allowing for the identification of unique behavioral profiles among individuals in the dataset.

B. Comparison of Methodologies

In this regard, Table IX shows the main differences between the most widely used data mining methodologies: KDD, SEMMA, and CRISP-DM. While KDD focuses on knowledge discovery from large volumes of data without a strict methodological structure, SEMMA offers a practical approach but is limited to the use of proprietary tools such as SAS and strictly technical phases. In contrast, CRISP-DM stands out for its comprehensive, iterative, and flexible nature,



(a) Silhouette score evaluation.

(b) Elbow method (Inertia).

(c) Cluster visualization (k=8).

Fig. 5. Evaluation and visualization of k-means clusters.

TABLE VIII. Cluster Analysis Results

(A) Metrics per k

k	Inertia	Silhouette
2	3245.03	0.18
3	2800.44	0.17
4	2416.18	0.19
5	2139.13	0.19
6	1899.30	0.20
7	1746.23	0.21
8	1563.09	0.22
9	1451.26	0.22
10	1363.39	0.21

(B) Distribution

Cluster	Count	Percentage (%)
0	155	15.50
1	112	11.20
2	133	13.30
3	125	12.50
4	149	14.90
5	105	10.50
6	114	11.40
7	107	10.70

(C) Profiles (Averages)

Cluster	Age	Screen	Dream	Activity
0	24.06	4.38	6.92	4.82
1	19.87	8.83	8.04	2.12
2	20.71	9.83	7.61	7.83
3	23.86	8.88	5.35	2.95
4	20.03	6.30	4.88	8.20
5	17.59	4.32	7.89	6.38
6	17.80	3.86	5.72	2.45
7	16.72	9.28	5.51	4.22

covering everything from business understanding to model implementation, making it a robust and widely validated standard in the industry. For these reasons, CRISP-DM is selected as the most appropriate methodology for this project, as it guarantees complete and systematic coverage of the data mining process.

V. Discussion

The study described by [14] demonstrated that unsupervised clustering algorithms make it possible to segment the mental health of university students into clearly differentiated profiles, particularly when employing $k = 3$, supported by the silhouette index. However, unlike that work, which integrated classification and clustering into a hybrid model to optimize the prediction of academic performance and psychological well-being, the present research focuses on a purely unsupervised exploratory approach. This approach, by dispensing with prior labels, facilitates the identification of hidden patterns and enables the comparison of how adjusting the number of clusters ($k = 3$ versus $k = 8$) impacts the balance between clarity and level of detail in segmentation. In doing so, it highlights the usefulness of pure clustering for generating initial analytical foundations that could complement hybrid models in future studies [15].

Complementarily, the work of [16] highlights the positive impact of structured interventions on improving psychological well-being, employing a Bayesian model and cluster sampling to evaluate dimensions such as mental health literacy, self-efficacy, and willingness to seek help. The findings of the present study align with these results by showing that segmentation strategies promote the identification of vulnerable groups and, consequently, the strengthening of such dimensions, albeit with nuances related to the academic and social context of the students analyzed. In this sense, it is evident that, regardless of the methodology applied, intervention programs generate consistent effects in promoting student well-being.

On the other hand, the research of [17] introduces an approach based on wearable devices and physiological signals (EEG, EDA, HR, and SKT), achieving accuracy levels above 99% in stress classification in controlled environments using the MIST. While these results demonstrate the effectiveness of real-time monitoring, they present limitations regarding their applicability in everyday scenarios. In contrast to this approach, the results obtained here prioritize contextual variables such as lifestyle and academic habits, which, although yielding lower levels of accuracy, provide a broader and more representative perspective of the students' real environment. Thus, a contrast is established between the high accuracy of wearables under experimental conditions and the practical applicability of contextual factors in real educational settings.

Likewise, compared with studies that employed questionnaires and classical classification algorithms such as Naïve Bayes, Decision Trees, Random Forest, and SVM, achieving accuracies between 71% and 78.9% for predicting anxiety levels [18] [19], the results obtained in the present study provide an alternative approach by integrating clustering techniques and contextual analysis. Although supervised algorithms allow direct predictions of specific symptoms, clustering offers a complementary perspective by identifying group patterns that transcend individual characteristics, thus providing a more solid foundation for preventive interventions of a collective nature.

In the same vein, studies such as those of [20] [21] employed supervised models including multiple linear regression, random forest, and multilayer neural networks to predict depression, suicidal ideation, and stress levels, highlighting specific factors such as financial concerns or somatic symptoms as key predictors. In contrast, the present study emphasizes group analysis through clustering, revealing how students can be organized into profiles of shared well-being and risk. This perspective makes it possible to move beyond the prediction of individual cases and focus instead on the understanding of collective trends that can guide broader preventive interventions tailored to different segments of the student population.

When contrasting the findings of Rois [22] and Saha et al. [23] with the results of this study, a convergence emerges regarding the relevance of external factors in explaining student mental well-being. While those authors identified physiological variables such as heart rate, blood pressure, or sleep quality using supervised algorithms—highlighting random forest with an accuracy of 89.72% and an AUC of 0.8715—as well as the utility of social media for anticipating mental health consultations with a correlation of $r = 0.86$, this work applied unsupervised clustering techniques that allowed students to be segmented into groups with similar risk and mental prosperity characteristics. Unlike predictive approaches based on supervised models or large-scale data analysis, the results presented here offer an exploratory perspective that reveals collective patterns without requiring prior labels, complementing previous findings by facilitating a comprehensive understanding of how academic, personal, and contextual factors combine into differentiated mental health profiles.

Similarly, the studies by Abdul [24] [25] used physiological variables, lifestyle habits, and large-scale social media data, supported by algorithms such as random forest, SVM, and time series models, achieving remarkable performance metrics (AUC of 0.8715 and correlations up to $r = 0.86$). However, while those approaches privilege individual prediction, the results presented here underscore the capacity of clustering to provide a global view of how students are distributed across different profiles of risk or well-being. Thus, a methodological complementarity is proposed: supervised models are useful for anticipating individual cases, while unsupervised techniques allow for the understanding of collective structures that strengthen the design of prevention strategies and comprehensive monitoring.

Finally, the findings are consistent with those reported by [26] [27], who demonstrated the usefulness of clustering in segmenting university mental health profiles based on factors such as alcohol consumption, perceived stress, or academic cohorts. While these authors applied k-means and other clustering methods with specific variables, the present study jointly integrates academic and personal factors, expanding the scope of analysis and providing a more holistic perspective of student well-being. Taken together, the results reaffirm that

TABLE IX. Comparison of Data Mining Methodologies

Criterion	KDD	SEMMA	CRISP-DM
Main Focus	Knowledge discovery from large volumes of data	Analytical process focused on statistical modeling	Standardized data mining process applied to different domains
Methodological Structure	Does not have rigidly defined phases, more conceptual	Technical phases: Sample, Explore, Modify, Model, Assess	Clear phases: Business Understanding, Data Understanding, Preparation, Modeling, Evaluation, Deployment
Flexibility	Low flexibility; requires adaptation to context	Limited to projects using SAS software	High flexibility and iterativity, applicable to diverse contexts
Scope	More theoretical discovery than practical implementation	Focused on exploratory analysis and model building	Covers the entire lifecycle: from business to implementation
Advantages	Pioneer in formalizing the process of pattern discovery	Direct and practical in SAS projects	Industry standard, robust, documented, and widely used
Limitations	Lacks standardization for real projects	Dependence on proprietary software, not very adaptable to other environments	May be more extensive and complex compared to simpler methodologies

clustering is not only effective for classification but also for generating segmentations that enable the design of timely and tailored interventions for differentiated profiles within the university population.

VI. Conclusion

Mental health in the university setting has shown various issues that significantly affect the academic and personal lives of students. These difficulties not only influence academic performance but also impact behavior, interpersonal relationships, and emotional well-being. In this context, the research conducted proposed the creation of a predictive model based on machine learning, using the k-means algorithm as the main approach. To develop the proposed model, a dataset in CSV format, oriented toward university students' mental health, was extracted from the Kaggle platform. The research was structured following the CRISP-DM methodology, which establishes five stages in its development. The first stage refers to business understanding based on the identification of the problem. As a second stage, an analysis of the most relevant variables from the dataset was conducted, applying criteria such as mean, maximum, minimum, among others. The third stage involved analyzing outliers, performing data cleaning, and transforming categorical variables into numerical variables for further evaluation. In the fourth stage, the k-means algorithm was applied to cluster the data and evaluate the optimal number of groups created. Finally, the fifth stage provides an evaluation of metrics based on the established clusters.

The results obtained confirm that cluster analysis through PCA, along with metrics such as Silhouette and inertia, makes it possible to identify significant patterns within the student population. It is observed that configurations such as $k = 3$ and $k = 8$ generate consistent groupings, differentiated by key variables such as screen time, hours of sleep, and level of physical activity. These segmentations not only facilitate the understanding of students' distribution and behavior in the academic environment but also help identify subgroups with specific needs or habits. Altogether, this approach provides relevant information for designing academic support strategies and programs aimed at well-being and mental health, contributing to a deeper understanding of the factors affecting university performance and adaptation.

Based on these findings, it is recommended that future research incorporate additional variables that allow for a more comprehensive characterization of students, such as eating habits, use of social networks, socioeconomic level, and resilience. Likewise, it is advisable to complement cluster analysis with different algorithms such as DBSCAN, Agglomerative Clustering, or Gaussian Mixture Models, as well as to explore advanced dimensionality reduction techniques such as t-SNE or UMAP to improve data visualization. It is also suggested to evaluate cluster stability through cross-validation or bootstrap techniques, integrate temporal data to analyze changes over time, and combine clustering with supervised predictive models to anticipate academic or mental health risks. Finally, complementing the study with interactive tools or dashboards would facilitate the interpretation and application of results, strengthening the model's usefulness for decision-making in educational environments.

The developed model presents certain limitations that must be considered when interpreting the results. First, it is based solely on the variables available in the original dataset; therefore, relevant external factors for mental health and academic performance may not be reflected. Moreover, since this is an unsupervised cluster analysis, the assignment of students to groups depends on the algorithm configuration and the choice of the number of clusters, which may generate some variability in the results. Another limitation is that the model does not incorporate temporal data, so it does not allow for the evaluation of how students' behavioral patterns change over time. Lastly, although quality metrics such as Silhouette and inertia help evaluate cluster cohesion and separation, they do not guarantee an absolute interpretation of the practical relevance of each group, making it necessary to complement the analysis with input from education and mental health experts.

Acknowledgment

Thanks to the University of Sciences and Humanities, which made this project a reality.

References

[1] K. Wantala, F. C. L. S. Go, V. C. C. Garcia, P. Chirawatkul, N. Chanlek, P. Kidkhunthod, R. R. M. Abarca, and M. D. G.

de Luna, "Low thermal oxidation of gaseous toluene over cu/ce single-doped and co-doped oms-2 on different synthetic routes," *Chemical Engineering Communications*, vol. 211, pp. 350–365, 2024, doi:10.1080/00986445.2022.2050710.

[2] D. A. Doubblestein, B. A. Spinelli, A. Goldberg, C. A. Larson, and A. M. Yorke, "Use of outcome measures by certified lymphedema therapists with survivors of breast cancer with breast cancer-related lymphedema," *Rehabilitation Oncology*, vol. 41, pp. 34–46, 1 2023, doi:10.1097/01.REO.0000000000000310.

[3] E. K. Ponce, M. F. Cruz, and L. Andrade-Arenas, "Machine learning applied to prevention and mental health care in peru," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 823 – 831, 2022, doi:10.14569/IJACSA.2022.0130196.

[4] C. Zheng, S. Liu, L. Song, Z. Xu, J. Guo, Y. Ma, Q. Ju, and J. Wang, "Comparison of sensible and latent heat fluxes from optical-microwave scintillometers and eddy covariance systems with respect to surface energy balance closure," *Agricultural and Forest Meteorology*, vol. 331, 3 2023, doi:10.1016/J.AGRFORMET.2023.109345.

[5] J. A. Harris, P. K. Guntaka, C. J. Niedziela, S. R. Aziz, and S. Afshar, "Interest in global surgery rotations among oral and maxillofacial surgical residents in the united states," *Journal of Dental Education*, vol. 88, pp. 30–41, 1 2024, doi:10.1002/JDD.13394.

[6] C. O. M. Arroyave, D. C. Arango, D. A. Restrepo-Ochoa, and A. C. Calvo, "Positive mental health: between well-being and the development of capabilities," *CES Psicología*, vol. 15, no. 2, 2022, doi:10.21615/cesp.5275.

[7] J. Soihet and A. D. Silva, "Psychological and metabolic effects of food restriction in binge eating disorder," *Nutrição Brasil*, vol. 18, no. 1, 2019, doi:10.33233/nb.v18i1.2563.

[8] M. S. Suzigan, L. da Ressurreição Santos, S. A. Rosa, E. R. Caldas, B. M. D. Araújo, F. de Castro Caixeta, A. A. D. Silva, and M. V. Ferreira, "Neurobiology of anxiety disorders," *Brazilian Journal of Health Review*, vol. 7, no. 1, 2024, doi:10.34119/bjhrv7n1-492.

[9] S. S. Valente, A. V. Padoin, D. S. Valente, C. L. de Sousa Brito, C. C. Mottin, and L. B. Micheletto, "Impact of psychological factors on bariatric surgery failure," *Psico*, vol. 54, no. 1, 2023, doi:10.15448/1980-8623.22.1.39907.

[10] C. Prado, M. Santero, D. Caruso, F. Ortiz, M. S. Zamorano, and V. Irazola, "What are the knowledge gaps and research priorities in mental health of older adults? a mixed-method study using the combined strategies matrix for argentina (meca)," *Global Health Promotion*, vol. 30, pp. 87–94, Mar. 2023, doi:10.1177/17579759221086282.

[11] N. O. Hidayati, E. Widianti, D. A. Amira, Alfiatullatifah, R. H. Pratama, and R. R. N. Asifa, "Elderly in prison: A scoping review of mental health problems," *Enfermeria Global*, vol. 23, pp. 503–513, 2024, doi:10.6018/EGLOBAL.563741.

[12] M. Y. de Fatima Cucho Hidalgo and D. L. J. G. Lola, "Operational management of mental health services for the development of psychological well-being in post-pandemic university students," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 7, pp. 1284–1301, 1 2023, doi:10.37811/CL_RCM.V7I1.4481.

[13] A. Gonzalez-Mora and M. Gomez-Vargas, "Social relationships and mental health of migrant mothers and their children in medellin, colombia," *Revista Virtual Universidad Catolica del Norte*, pp. 140–167, Jan. 2023, doi:10.35575/RVUCN.N68A7.

[14] H. Singh and B. Kaur, "Predicting student performance and its impact on mental health using machine learning," *SSRN Electronic Journal*, 2023, doi:10.2139/ssrn.4356166.

[15] J. Gera, E. B. Marwaha, R. Thareja, R. Thareja, S. Gupta, and A. Jain, "Analysing student's academic performance in relation to psychosocial aspects using ai," *International Journal of Engineering Trends and Technology*, vol. 72, no. 1, 2024, doi:10.14445/22315381/IJETT-V72I1P124.

[16] X. Li, "Short-term rainfall monitoring in mountainous area based on bayesian model and mental health intervention of college students," *Arabian Journal of Geosciences*, vol. 14, no. 16, 2021, doi:10.1007/s12517-021-08010-5.

[17] V. Chandra and D. Sethia, "Machine learning-based stress classification system using wearable sensor devices," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 337–347, 2024, doi:10.11591/ijai.v13.i1.pp337-347.

[18] F. Fang, W. Chung, C. Ventre, M. Basios, L. Kanthan, L. Li, and F. Wu, "Ascertaining price formation in cryptocurrency markets with machine learning," *European Journal of Finance*, vol. 30, no. 1, p. 78–100, 2024, doi:10.1080/1351847X.2021.1908390.

[19] S. S. Rajkishan, A. J. Meitei, and A. Singh, "Role of ai/ml in the study of mental health problems of the students: a bibliometric study," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 5, pp. 5639–5654, 2024, doi:10.1007/s13198-023-02052-6.

[20] N. Meda, S. Pardini, P. Rigobello, F. Visioli, and C. Novara, "Frequency and machine learning predictors of severe depressive symptoms and suicidal ideation among university students," *Epidemiology and Psychiatric Sciences*, vol. 32, p. e55, 2023, doi:10.1017/S2045796023000550.

[21] I. J. Ratul, M. M. Nishat, F. Faisal, S. Sultana, A. Ahmed, and M. A. A. Mamun, "Analyzing perceived psychological and social stress of university students: A machine learning approach," *Heliyon*, vol. 9, no. 6, p. e17307, 2023, doi:10.1016/j.heliyon.2023.e17307.

[22] R. Rois, M. Ray, A. Rahman, and S. K. Roy, "Prevalence and predicting factors of perceived stress among bangladeshi university students using machine learning algorithms," *Journal of Health, Population and Nutrition*, vol. 40, no. 1, p. 50, 2021, doi:10.1186/s41043-021-00276-5.

[23] K. Saha, A. Yousuf, R. L. Boyd, J. W. Pennebaker, and M. D. Choudhury, "Social media discussions predict mental health consultations on college campuses," *Scientific Reports*, vol. 12, no. 1, p. 17928, 2022, doi:10.1038/s41598-021-03423-4.

[24] H. A. Rahman, M. Kwicklis, M. Ottom, A. Amornsriwatanakul, K. H. Abdul-Mumin, M. Rosenberg, and I. D. Dinov, "Machine learning-based prediction of mental well-being using health behavior data from university students," *Bioengineering*, vol. 10, no. 5, p. 575, 2023, doi:10.3390/bioengineering10050575.

[25] Y. Saxena, A. K. Mishra, D. Arora, and R. Devi, "Emotion based mental health classifier for ncr based engineering students," in *Proceedings of the 6th International Conference on Contemporary Computing and Informatics (IC3I 2023)*. IEEE, 2023, pp. 285–290, doi:10.1109/IC3I59117.2023.10397910.

[26] S. Lee, J. Lim, S. Lee, Y. Heo, and D. Jung, "Group-tailored feedback on online mental health screening for university students: using cluster analysis," *BMC Primary Care*, vol. 23, no. 1, p. 33, 2022, doi:10.1186/s12875-021-01622-6.

[27] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, pp. 24 153–24 185, 2024, doi:10.1007/s11042-023-16407-5.

[28] S. Z. E. Mestari, G. Lenzini, and H. Demirci, "Preserving data privacy in machine learning systems," *Computers and Security*, vol. 137, 2024, doi:10.1016/j.cose.2023.103605.

[29] D. Peral-García, J. Cruz-Benito, and F. J. García-Peñalvo, "Systematic literature review: Quantum machine learning and its applications," *Computer Science Review*, vol. 51, 2024, doi:10.1016/j.cosrev.2024.100619.

[30] J. J. E. Zúñiga, "Application of crisp-dm methodology for geographic segmentation of a public database," *Ingeniería Investigación y Tecnología*, vol. 21, no. 1, pp. 1–15, 2020, doi:10.22201/fi.25940732e.2020.21n1.008.

[31] R. Clancy, D. O'Sullivan, and K. Bruton, "Data-driven quality improvement approach to reducing waste in manufacturing," *TQM Journal*, vol. 35, no. 1, pp. 51–72, 2023, doi:10.1108/TQM-02-2021-0061.

[32] C. E. D. Vanegas, J. C. G. Mejía, F. A. V. Agudelo, and D. E. S. Duran, "A representation based on essence for the crisp-dm methodology," *Computacion y Sistemas*, vol. 27, no. 3, pp. 675–689, 2023, doi:10.13053/CYS-27-3-3446.

[33] J. Brzozowska, J. Pizoń, G. Baytikenova, A. Gola, A. Zakimova, and K. Piotrowska, "Data engineering in crisp-dm process production data – case study," *Applied Computer Science*, vol. 19, no. 3, pp. 83–95, 2023, doi:10.35784/ACS-2023-26.

[34] J. Bokrantz, M. Subramaniyan, and A. Skoogh, "Realising the promises of artificial intelligence in manufacturing by enhancing crisp-dm," *Production Planning and Control*, vol. 35, pp. 2234–2254, 2024, doi:10.1080/09537287.2023.2234882.

[35] V. Krishnaswamy, N. Singh, M. Sharma, N. Verma, and A. Verma, “Application of crisp-dm methodology for managing human-wildlife conflicts: An empirical case study in india,” *Journal of Environmental Planning and Management*, vol. 66, no. 11, pp. 2247–2273, 2023, doi:10.1080/09640568.2022.2070460.

[36] A. Pambudi, “Application of crisp-dm using mlr k-fold on stock data of pt. telkom indonesia (persero) tbk: A case study on the indonesia stock exchange 2015–2022,” *Jurnal Data Mining dan Sistem Informasi*, vol. 4, no. 1, pp. 1–15, 3 2023, doi:10.33365/JDMSI.V4I1.2462.

[37] V. Plotnikova, M. Dumas, and F. P. Milani, “Applying the crisp-dm data mining process in the financial services industry: Elicitation of adaptation requirements,” *Data and Knowledge Engineering*, vol. 139, p. 102013, 5 2022, doi:10.1016/J.DATAK.2022.102013.

[38] A. Cheng, “Evaluating fintech industry’s risks: A preliminary analysis based on crisp-dm framework,” *Finance Research Letters*, vol. 55, p. 103966, 7 2023, doi:10.1016/J.FRL.2023.103966.