Clustering Sleep Patterns and Health Metrics Using K-Means Algorithm to Identify Profiles of Sleep Quality and Well-being in a Diverse Population

Abednego Dwi Septiadi[1,*], Muhamad Awiet Wiedanto Prasetyo[2]

[1]*Software Engineering, Telkom University, Indonesia*

[2]*Information System, Telkom University, Indonesia*

(Received: January 5, 2025; Revised: February 12, 2025; Accepted: May 12, 2025; Available online: July 23, 2025)

Abstract

This study explores the relationship between sleep patterns, health metrics, and lifestyle factors using K-Means clustering to identify distinct profiles based on these variables. The dataset comprises 374 individuals and includes features such as sleep duration, sleep quality, physical activity, stress levels, heart rate, blood pressure, BMI category, and sleep disorders. Data preprocessing, including the encoding of categorical variables and splitting the 'Blood Pressure' column into systolic and diastolic values, was performed before clustering. The optimal number of clusters was determined using the Silhouette Score, resulting in 10 clusters. Each cluster was analyzed for its average characteristics, revealing that factors such as physical activity, stress, and sleep disorders significantly influenced sleep quality. The findings emphasize the importance of managing stress and maintaining an active lifestyle to improve sleep quality, while also highlighting the need for addressing sleep disorders in individuals with poor sleep quality. This research provides valuable insights for personalized health recommendations and future interventions aimed at improving sleep hygiene and overall well-being.

Keywords: Blood Pressure, Clustering, Health Metrics, K-Means, Sleep Quality

1. Introduction

Sleep quality is a critical element of overall health that has garnered significant attention in recent years, particularly regarding its severe impacts on both physical and mental well-being. A growing body of research underscores that insufficient or poor-quality sleep is associated with various adverse health outcomes, ranging from cognitive impairments to chronic diseases. Studies have demonstrated that the duration and quality of sleep are intricately linked to physical, social, and emotional health outcomes among various populations, including adults and children [1]. The crucial role sleep plays in regulating both physiological and psychological processes makes it an indispensable component of health maintenance. Research indicates that sleep disturbances can impede cognitive development in children, with implications for their growth and emotional regulation [2]. For instance, infants who achieve a higher quality of sleep demonstrate improved cognitive abilities, which can lead to better academic performance and emotional health throughout their lives [2]. These findings suggest that interventions designed to enhance sleep quality during infancy could yield substantial long-term health benefits. Furthermore, comparable patterns have been observed in adults, where good sleep quality significantly influences psychological factors such as emotion regulation and overall mental health [3].

The psychological ramifications of poor sleep quality extend beyond mere cognitive performance. Poor sleep has been linked to increased rates of anxiety, depression, and other mood disorders [4]. A meta-analysis revealed that improving sleep quality could ameliorate the symptoms of these mental health conditions [3]. Moreover, chronic sleep deprivation has been identified as a precursor to substance use disorders, suggesting a complex interplay between sleep, addiction, and mental health [5]. The challenges of maintaining sleep quality have been exacerbated by external stressors such as the COVID-19 pandemic, leading to increased anxiety and altered sleep patterns among various demographics,

***Corresponding author: Abednego Dwi Septiadi (abednego@telkomuniversity.ac.id)**

DOI: https://doi.org/10.47738/ijaim.v5i2.104

particularly students and healthcare workers [6]. Physical health is equally affected by sleep quality. Quality sleep is pivotal for metabolic regulation, immune function, and recovery from injuries. Poor sleep has been associated with a range of chronic health issues, including obesity, diabetes, and cardiovascular diseases [7]. For instance, individuals who experience poor sleep tend to exhibit higher dietary risks and poor lifestyle choices, leading to serious health complications [7]. The immunological impacts of sleep deprivation further elucidate its role in health, indicating that compromised sleep can weaken the immune system and elevate susceptibility to viral infections such as COVID-19 [8].

Clustering sleep patterns and health metrics serves as a pertinent approach to understanding the complex relationship between sleep quality and health outcomes. The main objective of such analyses is to identify diverse profiles that capture variations in sleep quality alongside associated health metrics, thereby facilitating targeted interventions and preventative strategies. By leveraging clustering frameworks, researchers can explore the multifaceted dynamics surrounding sleep, including its effects on physical and mental health, which has become increasingly critical in contemporary health discourse. Distinguishing different sleep profiles based on health metrics can illuminate how varying sleep patterns correlate with physical and mental health conditions. For instance, studies have demonstrated that inadequate sleep can lead to negative health outcomes, including depression, increased healthcare costs, and diminished quality of life [9], [10]. Investigating these correlations through clustering can help identify specific at-risk groups whose sleep disturbances may predispose them to further mental health issues [9]. By gathering and analyzing data on sleep quality, duration, hygiene, and associated health metrics, researchers aim to provide nuanced insights into how poor sleep hygiene exacerbates conditions such as anxiety and depression, while informing targeted treatment protocols [11].

Identifying profiles that correlate poor sleep with detrimental health outcomes underscores the necessity of addressing these issues within management frameworks. For example, individuals exhibiting insomnia or disrupted sleep patterns have reported exacerbated symptoms of PTSD, depression, and anxiety, suggesting that sleep quality is essential in mitigating these mental disorders [10]. Such findings support the notion that treating sleep disorders could yield additional benefits across a range of mental health conditions [10], highlighting the need for policymakers and health practitioners to incorporate sleep medicine into broader mental health strategies. Additionally, integrating lifestyle factors—such as physical activity and dietary habits—into clustering methodologies can offer a more comprehensive view of an individual's health profile concerning sleep quality. Evidence indicates that regular physical activity is associated with improved sleep quality and better overall health outcomes [12], [13]. By clustering individuals based on their physical activity levels alongside sleep quality, researchers can pinpoint specific demographics that may require interventions to promote exercise alongside improved sleep hygiene. This interplay between sleep and activity patterns highlights the potential for health strategies that enhance both physical and mental health through optimized sleep [14].

The comprehensive understanding of different sleep profiles is critical for enhancing sleep hygiene and improving overall well-being. The repercussions of poor sleep are evident in both physical and mental health domains, necessitating health researchers and practitioners to investigate the factors contributing to sleep quality. By elucidating the relationships between sleep profiles and health outcomes, health interventions can be tailored effectively to improve individual sleep hygiene practices. A fundamental aspect of sleep hygiene relates to behavioral practices that are conducive to consistent good sleep. Research indicates that individuals with a robust understanding of sleep hygiene demonstrate better quality sleep and experience fewer comorbidities related to sleep disorders [15]. For instance, college students who receive education regarding sleep hygiene principles—including the avoidance of stimulants like caffeine and regulation of their sleep environment—report higher sleep quality, which can lead to improved academic performance and reduced levels of anxiety and depression [16]. Therefore, implementing educational interventions focused on sleep hygiene can enhance awareness among vulnerable populations, such as college students, who often face stressors that compromise their sleep patterns.

The scope of this paper focuses on analyzing a dataset that includes various sleep and health metrics, such as sleep duration, sleep quality, physical activity levels, stress levels, BMI, blood pressure, heart rate, and daily steps. The dataset aims to explore how these factors relate to overall sleep quality and well-being. To achieve the research objective of identifying distinct profiles based on sleep patterns, the paper applies the K-Means clustering algorithm, which groups individuals with similar health and sleep characteristics. The method allows for the identification of meaningful clusters, which can provide insights into different sleep profiles and their associated health risks.

2. Literature Review

2.1. Sleep Quality and Health

Understanding the connection between sleep duration, sleep quality, and health metrics is critical in promoting holistic health and well-being. Numerous studies have highlighted that both inadequate sleep duration and poor-quality sleep are associated with various adverse health outcomes, thus underscoring the necessity of addressing both aspects in public health initiatives. The following analysis reviews pivotal research that elucidates these interactions. Research consistently demonstrates that sleep duration plays a crucial role in overall health. Chronic short sleep duration—defined as less than 7 hours per night—is linked to a variety of health issues, including cardiovascular disease, obesity, and depressive symptoms [17]. In their cross-sectional study, Nutakor et al found that individuals who do not meet the recommended 7–9 hours of sleep per night exhibit higher susceptibility to these health conditions. This highlights the role of chronic sleep deprivation not only as a risk factor but also potentially as a causal agent for these diseases, given the physiological processes affected by inadequate rest [17].

In addition to sleep duration, the quality of sleep is a vital metric for assessing sleep health. Poor sleep quality, characterized by frequent awakenings, restlessness, or inability to remain asleep, correlates with numerous health complaints. Chen et al conducted a population-based study revealing associations between poor sleep quality and various chronic diseases, including conditions like hypertension and diabetes [18]. This relationship signifies that improving sleep quality could yield healthcare cost reductions and enhance individual health metrics. Furthermore, Del-Valle-Soto et al have highlighted physiological changes during sleep, noting that the body enters a state of reduced heart rate and metabolic activity, indicating the importance of good quality sleep for sustaining bodily functions and health [19]. Their research affirms that optimal sleep quality, alongside adequate duration, is essential for recovering physical health and establishing well-being, making it imperative for health interventions to promote both aspects.

The impact of socio-demographic factors on sleep is another crucial consideration illustrated in recent studies. Hansen et al found socioeconomic disparities significantly affect sleep duration and quality among children, suggesting that targeting sleep hygiene education in lower socioeconomic groups could promote better sleep patterns and, subsequently, healthier developmental outcomes [20]. Understanding these socio-demographic influences is vital for developing tailored interventions that address specific community needs regarding sleep health. Kļaviņa-Makrecka et al indicated that insufficient sleep duration is associated with increased health complaints among adolescents, including increased sensitivity to pain and overall poorer health self-reports [21]. Their study emphasizes that sleep duration does not merely correlate with health outcomes, but actively contributes to individuals' health perception and quality of life. This provides a compelling argument for health professionals to prioritize sleep assessments within routine health screenings, especially among vulnerable populations.

2.2. Clustering in Health Data

Clustering algorithms, particularly K-Means, have gained prominence in health-related datasets where their utility in grouping individuals with similar traits can inform personalized healthcare strategies and enhance treatment efficacy. This discussion reviews several notable studies that have successfully employed K-Means and other clustering techniques, illustrating their effectiveness in various health contexts. The K-Means clustering algorithm is fundamentally a partitional clustering method that partitions data into distinct groups based on similarity, typically defined using Euclidean distance [22]. Its popularity stems from its straightforward implementation and computational efficiency, making it suitable for large health datasets [23].

One pertinent example is the study by Sari, which explored the application of K-Means clustering on tuberculosis patients in Ethiopia [24]. By analyzing spatial and health data, the study successfully identified different patient groups based on their clinical responses and healthcare needs. This clustering allowed for more targeted public health interventions and resource allocation, illustrating K-Means' effectiveness in summarizing complex health data into actionable insights. In another study, Setiyaji and Purnomo implemented K-Means clustering to analyze the distribution of health workers in Semarang City, providing critical insights for workforce planning in healthcare delivery [25]. The study highlighted how clustering techniques can be utilized to optimize health service provision by ensuring adequate staffing across various regions.

While K-Means serves as a robust starting point for clustering health data, researchers have also integrated it with other techniques to enhance its performance [26]. For instance, Sharannavar and K emphasized the performance analysis of clustering algorithms for dyslexia detection, including an improved K-Means setup [27]. Their method utilized the Elbow method for optimal cluster determination, illustrating how combining K-Means with advanced techniques can yield more precise clustering results. The work by Paramita also illustrates an innovative approach by combining K-Means with Fuzzy C-Means clustering and Principal Component Analysis (PCA) to improve classification accuracy in data classification tasks [28]. This approach underscores the potential of integrating K-Means with dimensionality reduction and fuzzy logic techniques to address complexities specific to health datasets, paving the way for enhanced diagnostic tools.

2.3. Health Metrics and Sleep Hygiene

In exploring the relationship between sleep quality, sleep hygiene, and health metrics, several key formulas and models are employed to quantify and assess these variables. This review highlights metrics such as the Pittsburgh Sleep Quality Index (PSQI) for sleep quality, Body Mass Index (BMI) for assessing weight-related health risks, and various psychological scales for measuring stress and physical activity. Each measure provides a unique lens for understanding the interplay between sleep and overall health. The PSQI is a widely used instrument to assess sleep quality over the previous month. It comprises seven components: sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleep medication, and daytime dysfunction. Each component is rated on a scale from 0 to 3, with a higher score indicating worse sleep quality. The total scores can range from 0 to 21, where scores above 5 indicate poor sleep quality [29]. This tool is valuable as it captures both subjective aspects of sleep quality and objective measures like duration and disturbances, making it a comprehensive assessment for researchers and clinicians [30].

BMI is a common metric for evaluating body weight concerning height, serving as an indirect measure of body fat. The formula for BMI is:

$$BMI = \frac{weight(kg)}{height(m)^2} \tag{1}$$

BMI categories include underweight (BMI < 18.5), normal weight (18.5 ≤ BMI < 24.9), overweight (25 ≤ BMI < 29.9), and obesity (BMI ≥ 30) [31]. Studies have shown that sleep duration and quality can significantly influence BMI. For example, Xu et al demonstrated a correlation between habitual sleep duration and BMI, illustrating how sleep metrics can identify at-risk individuals for obesity-related health issues [32]. Stress levels can be quantified using various psychological scales, with the Perceived Stress Scale (PSS) being one of the most validated measures. The PSS assesses how unpredictable, uncontrollable, and overloaded respondents find their lives [33]. The PSS contains ten items rated on a 5-point Likert scale, allowing for scores ranging from 0 to 40, with higher scores indicative of greater perceived stress. Research indicates that elevated stress levels can disrupt sleep patterns, creating a cycle detrimental to overall health [33].

3. Methodology

3.1. Data Collection

The dataset utilized for this analysis was collected from Kaggle, in a CSV format, encompassing various health and lifestyle metrics that influence sleep patterns. The dataset includes features such as sleep duration, quality of sleep, physical activity levels, stress levels, BMI category, blood pressure, heart rate, daily steps, and sleep disorders. Each of these features provides valuable insights into the factors that contribute to both sleep patterns and overall health. The dataset was initially loaded using pandas, a powerful library for data manipulation and analysis. After loading the dataset, an inspection was carried out to assess its overall structure, including the number of rows and columns, the data types of each column, and the presence of any missing values or anomalies. Summary statistics for numerical features were calculated to provide an overview of the central tendency and spread of the data. Descriptive statistics were computed, including measures like mean, standard deviation, minimum, and maximum values, to ensure that the data was suitable for further analysis. Additionally, the dataset was checked for duplicate rows, ensuring that the data used in the analysis was clean, unique, and free from redundancy. Missing values, especially in columns such as 'Sleep Disorder', were handled by filling them with 'None', since sleep disorder types were non-numerical and would not affect the clustering process. This approach allowed the dataset to remain intact while ensuring that no valuable data was excluded due to missing values.

3.2. Data Preprocessing and Feature Engineering

Data preprocessing is a crucial step in any machine learning pipeline, particularly when preparing a dataset for clustering. In this analysis, several preprocessing steps were implemented to ensure that the dataset was ready for K-Means clustering. The first step involved removing the 'Person ID' column, as it was a unique identifier for individuals and did not contribute to clustering. Identifiers are not relevant in clustering algorithms, which rely on finding patterns based on features, not on individual identities. The next significant preprocessing task was dealing with the 'Blood Pressure' column, which contained systolic and diastolic values in a single string. For better granularity, the 'Blood Pressure' column was split into two separate columns: 'Systolic' and 'Diastolic'. These new columns were cast into integer format for numerical analysis, as K-Means requires numerical inputs to compute distances between data points.

In addition to this, categorical features such as 'Gender', 'Occupation', 'BMI Category', and 'Sleep Disorder' were encoded using the LabelEncoder from scikit-learn. Label encoding was selected because K-Means clustering works with numerical data, and LabelEncoder transforms categorical labels into integers, making it compatible with the algorithm. Label encoding is particularly useful here because the categorical variables did not have an excessive number of categories, thus preventing a high-dimensional feature space that could have complicated the clustering process. For instance, the 'Occupation' feature, which contained a variety of job titles, was encoded into numerical values. This transformation ensured that the clustering process would treat categorical features appropriately while maintaining simplicity. The 'Sleep Disorder' column, which contained some missing values, was filled with the value 'None' to avoid missing data issues during the encoding. This preprocessing also ensured that the categorical variables were numerically encoded without creating overly sparse data, which is common when using one-hot encoding in high-dimensional datasets. Once preprocessing and feature engineering were complete, the dataset was examined again to confirm that all transformations had been correctly applied. The processed dataset was now in a state suitable for clustering, with all features numerically encoded and any missing values handled.

3.3. Exploratory Data Analysis (EDA)

EDA was performed to better understand the structure of the dataset and the relationships between different features. EDA plays a vital role in data analysis by helping identify patterns, trends, and anomalies in the data. The first step in the EDA process was to visualize the distribution of numerical features using histograms. This visualization provided insight into how different features, such as sleep duration, physical activity, heart rate, and daily steps, were distributed across the population. The histograms helped identify features with skewed distributions, which could suggest the need for further transformation or normalization. For example, sleep duration might exhibit a normal distribution, while physical activity could be heavily skewed depending on the population. After analyzing the numerical features, the categorical features were visualized using count plots. These plots showed the frequency of different categories, such as BMI categories, types of sleep disorders, and gender. This step helped to reveal the distribution of these categories, providing useful insights into how these features were distributed in the population. For instance, the distribution of 'BMI Category' could indicate whether the dataset contains more individuals in the normal weight or overweight category, which could have implications for sleep quality.

A crucial part of the EDA was creating a correlation heatmap to examine the relationships between numerical features. The heatmap displayed the correlation coefficients between different features, helping to identify whether variables like sleep quality and physical activity were positively or negatively correlated. For example, we might expect a negative correlation between stress levels and sleep quality, and a positive correlation between daily steps and physical activity levels. By exploring these correlations, we could gain a better understanding of which variables might influence each other. Additionally, a pair plot was created to visualize the relationships between key features such as sleep duration, quality of sleep, physical activity levels, and stress levels, with the data color-coded by sleep quality. This visualization allowed for a more detailed exploration of how these factors interact with each other and their relationship to sleep quality. Pair plots are especially useful for detecting potential clusters or patterns in the data and are an effective way to visually inspect the data before proceeding with clustering.

3.4. Finding the Optimal Number of Clusters (K)

Determining the optimal number of clusters (K) is a critical step in any clustering task, as the number of clusters directly impacts the results. To identify the best value for K, the Elbow Method and the Silhouette Score were used. The Elbow Method involves calculating the Within-Cluster Sum of Squares (WCSS) for different values of K and plotting the results. The WCSS measures how tightly the data points are clustered around the centroids, with a lower value indicating tighter clusters. The optimal number of clusters is typically identified at the "elbow" point in the plot, where the rate of decrease in WCSS slows significantly. This point suggests that increasing the number of clusters beyond this value does not lead to significant improvements in cluster compactness. In addition to the Elbow Method, the Silhouette Score was also computed for different values of K. The Silhouette Score measures how similar each data point is to its own cluster compared to other clusters. A higher silhouette score indicates that the clusters are well-separated and distinct. By examining both the Elbow Method and the Silhouette Score, the optimal number of clusters was selected, ensuring that the resulting clusters were meaningful and well-separated. The analysis showed that the Silhouette Score suggested the best value for K, ensuring that the clusters were both compact and clearly separated.

3.5. K-Means Clustering and Analysis

After determining the optimal number of clusters, K-Means clustering was applied to the preprocessed data. K-Means is a widely used clustering algorithm that works by iteratively assigning each data point to one of the K clusters based on the distance to the cluster centroids. The K-Means algorithm was applied after scaling the data using StandardScaler to ensure that all features contributed equally to the clustering process. StandardScaler was used to standardize the data, ensuring that each feature had a mean of 0 and a standard deviation of 1. This step was crucial because K-Means uses distance-based metrics to assign clusters, and features with larger scales could dominate the clustering process if not standardized. Once the data was scaled, K-Means clustering was performed, and the resulting cluster labels were added to the original dataset. The next step was to analyze the resulting clusters by calculating the mean values of each feature within each cluster. This provided a profile for each cluster, allowing for an understanding of the typical characteristics of individuals in each cluster. For example, one cluster might represent individuals with high stress levels and poor sleep quality, while another cluster might represent individuals with good sleep quality and regular physical activity.

Additionally, the distribution of categorical features, such as sleep disorder types, was examined across the clusters. This analysis helped identify how lifestyle factors such as sleep disorders and BMI categories were distributed within each cluster. For example, one cluster might predominantly consist of individuals with sleep apnea, while another cluster might have a higher proportion of individuals with no sleep disorders. To visualize the clustering results, PCA was applied to reduce the dimensionality of the data to two components. This enabled the visualization of the clusters in a two-dimensional scatter plot. The clusters were color-coded to show how well-separated they were in the reduced-dimensional space. This visualization helped assess the quality of the clustering process and provided a clear, interpretable view of the cluster separation. Further analysis was conducted by comparing key features such as sleep duration, sleep quality, physical activity levels, and stress levels across clusters using box plots. This provided a detailed comparison of how these features varied across different clusters, offering insights into the factors that influenced the formation of each group.

To ensure that the K-Means model could be reused for future analysis, the trained K-Means model and the StandardScaler were saved as checkpoints using the joblib library. These checkpoints allow for easy reloading of the model and scaler, making it possible to apply the existing clusters to new data without retraining the model. By saving the model and scaler, the results of the clustering analysis can be reused and applied to different datasets, ensuring reproducibility and extending the usefulness of the model for future research. In conclusion, the method described in this study combines data preprocessing, exploratory analysis, and clustering to provide a comprehensive approach to understanding sleep patterns and their relationship to health and lifestyle factors. By identifying distinct clusters based on these factors, the study provides valuable insights into how sleep quality is influenced by various health and lifestyle metrics, helping to inform future interventions aimed at improving sleep hygiene and overall well-being.

4. Results and Discussion

4.1. Results of Data Inspection and Overview

The dataset was successfully loaded from the 'Sleep_health_and_lifestyle_dataset.csv' file, containing 374 entries and 13 columns. The first few rows of the dataset revealed that the key variables included demographic details (e.g., Gender, Age, Occupation), sleep metrics (e.g., Sleep Duration, Quality of Sleep), and health-related factors (e.g., Physical Activity Level, Stress Level, Heart Rate, Blood Pressure, Sleep Disorder). The dataset also showed that some features, such as 'Sleep Disorder', had missing values (219 missing entries), while other columns had no missing values. Notably, the dataset did not contain any duplicate rows, ensuring data integrity for the analysis. The summary statistics provided insights into the central tendencies of numerical columns. For example, the average sleep duration was approximately 7.13 hours, with sleep quality averaging 7.31 on a 1-9 scale. Physical activity levels were generally high, with an average activity level of 59, and the mean stress level was 5.39. Heart rate data averaged 70.17 beats per minute, and daily steps ranged from 3,000 to 10,000, with an average of 6,816 steps. The BMI category distribution was skewed toward individuals categorized as 'Normal' or 'Overweight', which might reflect broader population trends. The next step involved preprocessing the data to handle missing values and categorical features.

4.2. Data Preprocessing and EDA Finding

Data preprocessing involved several key transformations. The 'Person ID' column, being an identifier, was removed from the dataset. The 'Blood Pressure' column, which contained systolic and diastolic values in a single string format, was split into two separate columns—'Systolic' and 'Diastolic'. This split allowed for better numerical analysis, especially in relation to heart health. Categorical columns like 'Gender', 'Occupation', 'BMI Category', and 'Sleep Disorder' were encoded using LabelEncoder. This step converted non-numeric data into numeric form, making it compatible with the K-Means algorithm, which requires numerical data for clustering. For 'Sleep Disorder', missing values were filled with 'None', allowing the analysis to continue without imputation that could skew the results. After preprocessing, the dataset was well-structured for further analysis, and the transformed dataset had 374 rows and 13 columns, including the new 'Systolic' and 'Diastolic' columns and the encoded categorical variables.

EDA was performed to explore the data's underlying patterns and relationships. The distribution of numerical features was visualized using histograms (figure 1). Key findings include a relatively normal distribution for sleep duration, which averaged 7.13 hours, and a skewed distribution for physical activity, indicating that a large portion of the individuals were either sedentary or had a high level of activity. Stress levels showed moderate variation, with a higher proportion of individuals reporting stress levels between 4 and 6.

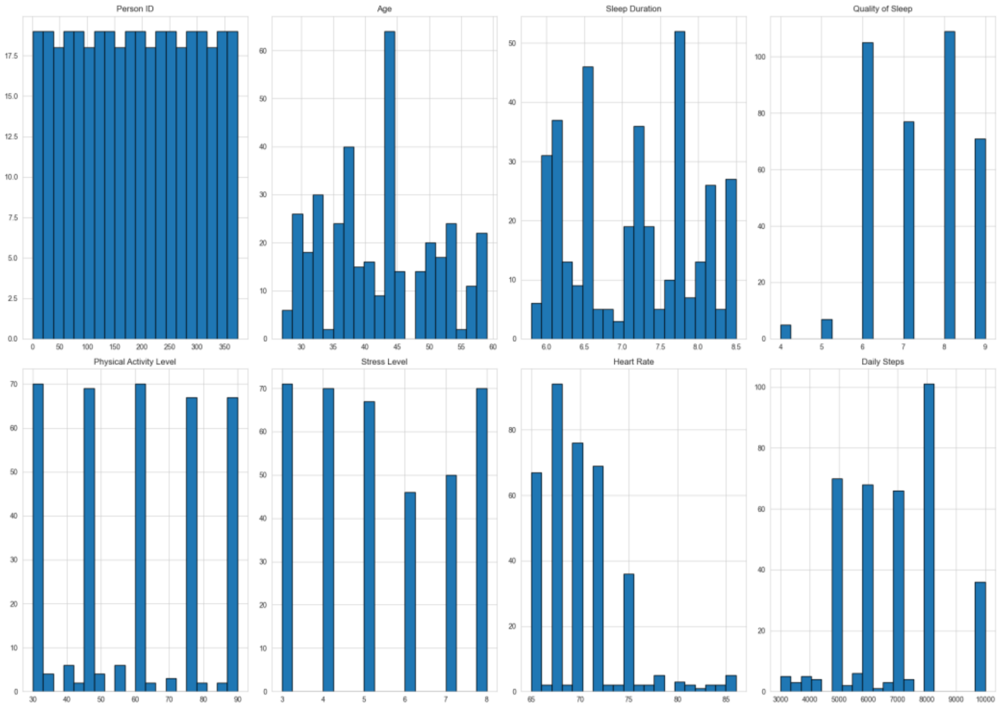

Figure 1. Distribution of Numerical Features

The correlation heatmap (figure 2) provided insight into relationships between numerical features. For example, sleep duration showed a moderate positive correlation with sleep quality, as expected. Other notable correlations included physical activity, which had a positive correlation with daily steps, suggesting that individuals with higher activity levels were likely to take more steps.

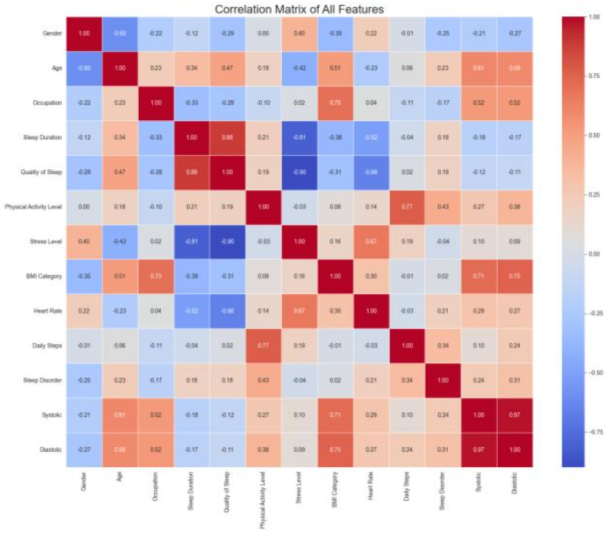

Figure 2. Correlation Matrix of All Features

A pair plot (figure 3) was also created to visualize relationships between key variables, such as sleep duration, quality of sleep, physical activity, and stress levels. This visualization helped to identify potential clusters and relationships between these variables.

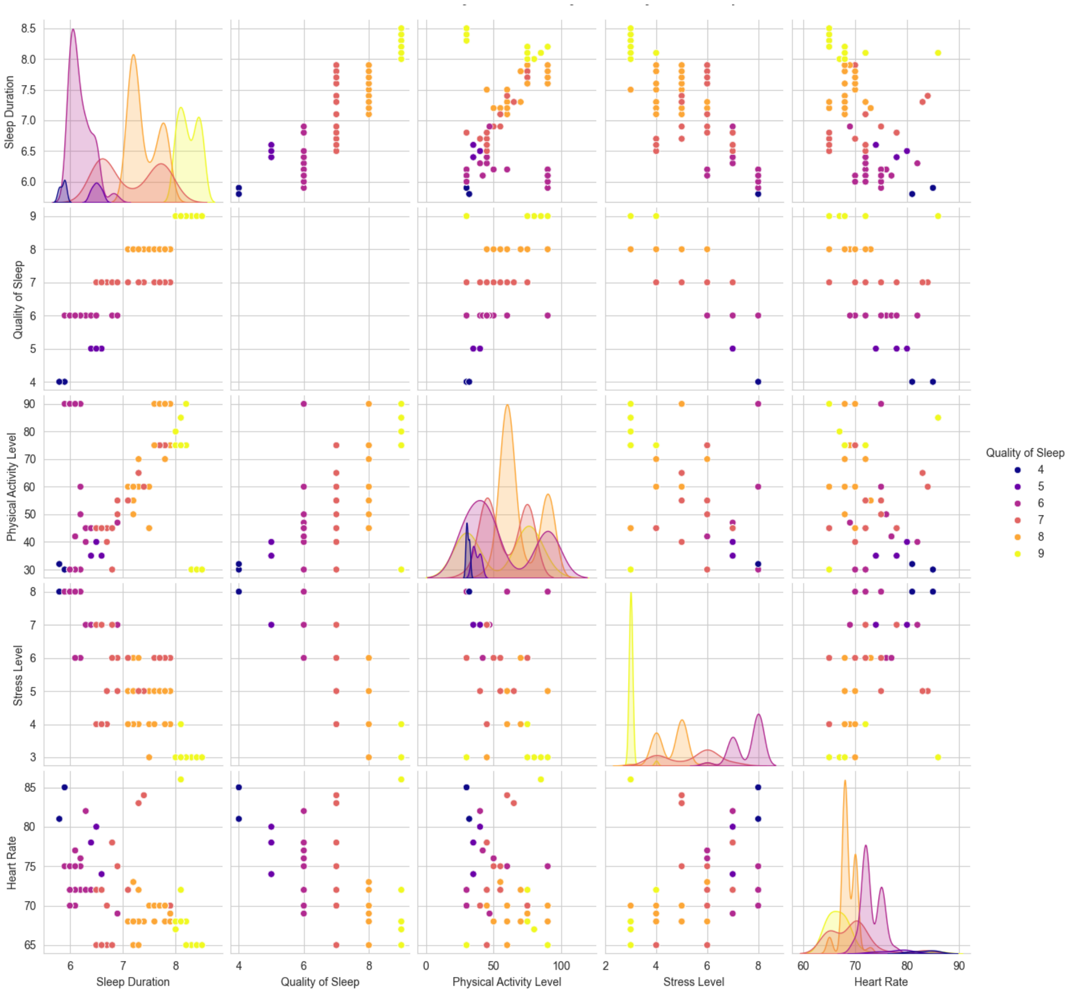

Figure 3. Pair Plot of Key Variables

4.3. K-Means Clustering and Cluster Profiling Results

The optimal number of clusters (K) for K-Means was determined using both the Elbow Method and the Silhouette Score. The Elbow Method suggested that an "elbow" point occurred between K=4 and K=5, where the rate of decrease in WCSS slowed down. However, based on the Silhouette Score, which evaluates the quality of the clusters, the optimal number of clusters was determined to be 10. The Silhouette Score method was preferred for its ability to provide a more accurate measure of cluster separation, and this value was selected for further analysis. The plot for both method shown in figure 4.

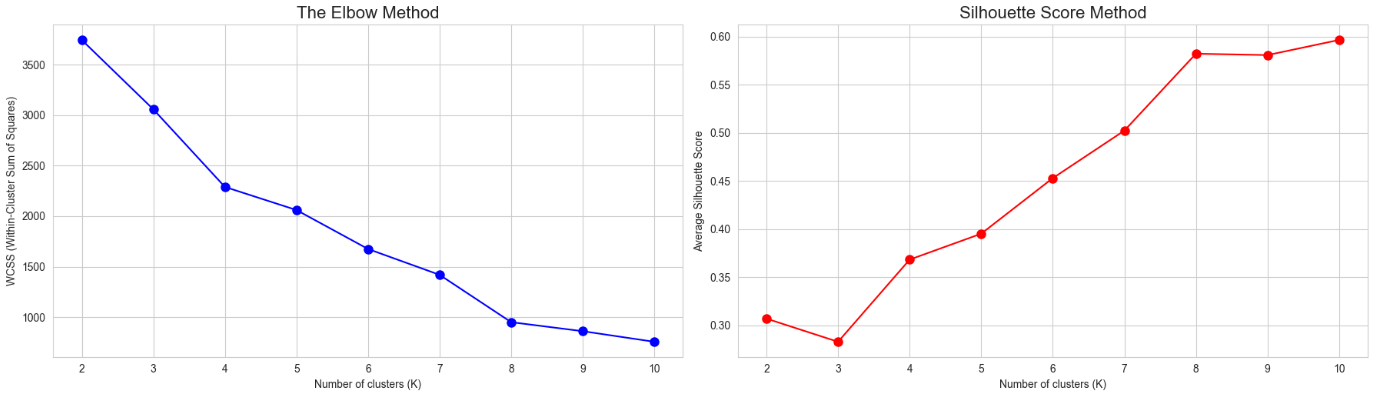

Figure 4. Determining Optimal K for Clustering with Elbow and Silhouette Score Method

With K=10 as the optimal number of clusters, K-Means clustering was applied to the preprocessed data. After scaling the data using StandardScaler, the clustering model was fit, and the resulting cluster labels were added to the dataset. The cluster profiles were analyzed by calculating the mean values of each feature within each cluster. The following key observations emerged from the cluster profiles. Cluster 0 had a relatively young age (mean age of 37.95 years), with good sleep quality (mean score of 8.10) and moderate physical activity levels (mean of 62.62). This cluster had a balance of moderate sleep duration (mean of 7.28 hours) and low stress levels (mean of 4.00). Cluster 1 comprised of older individuals (mean age of 58.03 years) with excellent sleep quality (mean of 9.00) and high physical activity levels (mean of 75.00). This cluster had the highest average sleep duration (8.09 hours) and the lowest stress levels (mean of 3.06). Cluster 2 had moderate sleep quality (mean of 6.05) and lower physical activity (mean of 44.23), coupled with high stress levels (mean of 6.82). The average sleep duration for this group was 6.46 hours.

Cluster 3 represented individuals with the lowest sleep quality (mean of 6.00) and high physical activity levels (mean of 90.00). They also had high stress levels (mean of 8.00). Cluster 4 were relatively young (mean age of 31.96 years) with good sleep quality (mean of 7.22) and moderate physical activity (mean of 69.33). Their stress levels were moderate (mean of 5.61). Cluster 5 had moderate sleep quality (mean of 7.00) and relatively low physical activity (mean of 43.89). Their stress levels were moderate (mean of 4.15), and they had an average sleep duration of 6.60 hours.

Cluster 6 displayed the lowest sleep quality (mean of 5.83) and moderate physical activity (mean of 30.43). Stress levels were high (mean of 7.94), and the sleep duration was relatively low (6.07 hours). Cluster 7 had high sleep quality (mean of 9.00) with low physical activity (mean of 30.00) and low stress levels (mean of 3.00). Their sleep duration was high (mean of 8.43 hours). Cluster 8 had good sleep quality (mean of 8.00) and high physical activity (mean of 77.03). Stress levels were moderate (mean of 4.98), and the sleep duration was 7.53 hours. Cluster 9 represented individuals with moderate sleep quality (mean of 7.67) and very low physical activity (mean of 70.00). Their stress levels were moderate (mean of 4.33), and the sleep duration was relatively high (7.60 hours).

4.4. Visualization of Clusters

The results of the K-Means clustering were visualized using PCA to reduce the data's dimensionality to two components. A scatter plot was created, showing the separation of clusters in a 2D space, with each cluster color-coded for easy identification (figure 5). The PCA plot clearly demonstrated that the clusters were well-separated, indicating that the K-Means algorithm had successfully identified distinct groups based on the sleep and health metrics. Additionally, box plots comparing key features such as sleep duration, sleep quality, physical activity, and stress levels across clusters were created (figure 6), providing a visual representation of how these features varied between clusters. These visualizations highlighted the differences in sleep quality and other metrics across the identified clusters.

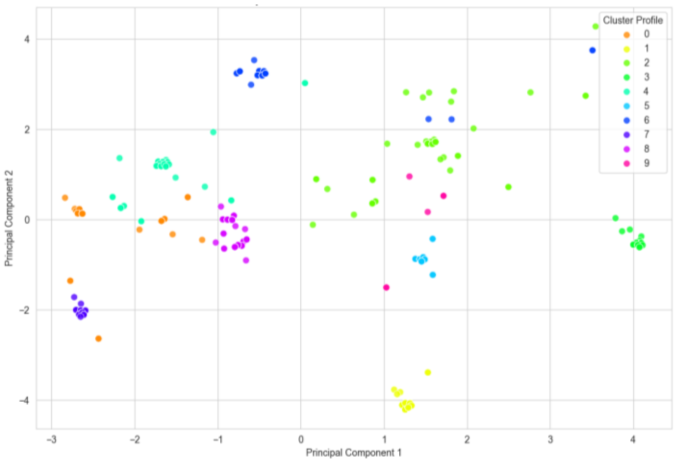

Figure 5. 2D PCA of Sleep & Health Data in 10 Cluster

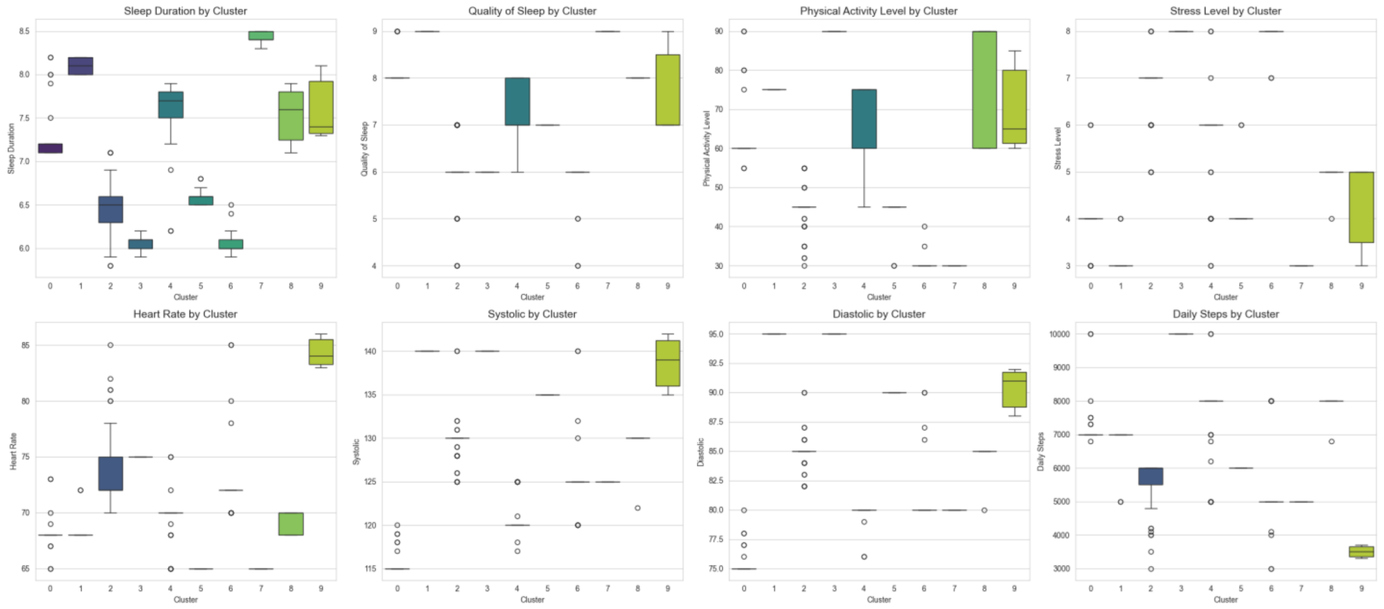

Figure 6. Comparison of Key Health Metrics across 10 Clusters

4.5. Discussion

The clustering analysis revealed significant variations in sleep quality and well-being across the 10 identified clusters, with key lifestyle and health factors playing a pivotal role. In Cluster 1, individuals exhibited excellent sleep quality, coupled with high physical activity levels and low stress. These individuals had the highest average sleep duration and the best overall sleep quality, suggesting that a balanced lifestyle with regular physical activity and low stress levels contributes positively to sleep quality. This finding highlights the importance of maintaining a healthy, active lifestyle for optimal sleep, supporting the notion that physical activity and stress management are crucial for good sleep hygiene. In contrast, Cluster 2, which displayed the lowest sleep quality, had individuals with high stress levels and lower physical activity. Despite their high physical activity levels, the significant presence of high stress contributed to the poor sleep quality observed in this group. This insight emphasizes that while exercise is beneficial, it may not be enough to ensure good sleep if stress levels are not adequately managed. Therefore, addressing mental well-being, alongside physical health, could be essential for improving sleep quality in individuals facing high stress. Another interesting finding emerged in Cluster 7, where individuals had very high sleep quality despite low physical activity and low stress levels. This suggests that factors beyond just physical activity, such as genetic predispositions, sleep hygiene habits, and possibly other unmeasured factors, might play a crucial role in sleep quality. While the analysis showed that individuals with good sleep quality tended to have higher activity levels, this cluster illustrates that other factors could also significantly contribute to restful sleep, emphasizing the complexity of sleep and well-being interactions. Lastly, the analysis of sleep disorders revealed that clusters with high proportions of individuals reporting conditions like 'Sleep Apnea' exhibited poorer sleep quality. For example, Cluster 3 and Cluster 6 had a significant proportion of individuals with sleep apnea, yet their stress levels and physical activity were not extreme. This finding underscores the impact of medical conditions on sleep quality, suggesting that individuals with sleep disorders may struggle to achieve high-quality sleep regardless of their physical or mental health status. Therefore, addressing underlying medical issues should be a priority in interventions aimed at improving sleep quality in populations with sleep disorders.

The clustering analysis identified 10 distinct profiles based on sleep patterns and health metrics, offering valuable insights into the complex relationship between lifestyle factors and sleep quality. The findings highlighted how physical activity, stress levels, and sleep disorders significantly influenced sleep quality. For instance, individuals with low stress and high physical activity generally exhibited better sleep quality, while those with high stress and sleep disorders, such as sleep apnea, struggled with poorer sleep quality. These profiles provide a nuanced understanding of how various factors, both physical and mental, interact to impact sleep and well-being. The practical implications of this research are far-reaching. The identified clusters can serve as a foundation for personalized sleep recommendations tailored to specific lifestyle and health profiles. By recognizing which factors most strongly influence sleep quality, healthcare providers can offer targeted advice to individuals struggling with sleep, whether it involves improving physical activity, managing stress, or addressing underlying sleep disorders. Additionally, public health interventions can be designed to address the broader population's sleep hygiene, focusing on lifestyle changes that promote better sleep, ultimately improving public health outcomes. Despite its valuable insights, the study has some limitations. The dataset used was relatively small and may not be fully representative of broader populations, which could limit the generalizability of the findings. Additionally, the lack of longitudinal data means that causal relationships between lifestyle factors and sleep quality cannot be conclusively established. Future research could address these limitations by expanding the dataset to include a more diverse sample and by collecting longitudinal data to observe how changes in lifestyle factors affect sleep quality over time. Furthermore, exploring more advanced clustering techniques, such as hierarchical clustering or deep learning-based clustering, could provide even more granular insights into sleep patterns and health. In conclusion, this study provides a valuable starting point for understanding the relationship between sleep and various health and lifestyle factors. While the findings offer useful insights for improving personalized sleep recommendations and public health interventions, further research is needed to validate these results and explore additional variables that may impact sleep quality. By addressing these gaps, future studies can enhance our understanding of how to optimize sleep for better overall health and well-being.

5. Conclusion

This clustering analysis successfully identified 10 distinct profiles from a dataset of 374 individuals, revealing the complex interplay between sleep patterns, health metrics, and lifestyle factors. The findings underscore that factors such as physical activity, stress levels, and the presence of sleep disorders significantly impact sleep quality. For instance, clusters with high physical activity and low stress generally reported better sleep quality, whereas high stress levels were linked to poorer sleep, even when physical activity was high. Furthermore, the presence of specific sleep disorders like sleep apnea was a strong determinant of poor sleep quality, sometimes irrespective of other lifestyle factors. These distinct profiles highlight that a multifactorial approach is necessary to understand and address sleep-related issues.

The practical implications of this research are significant, offering a foundation for creating personalized health recommendations and targeted public health interventions. By identifying an individual's cluster profile, healthcare providers can offer tailored advice, focusing on stress management, physical activity, or medical treatment for sleep disorders as needed. However, the study's limitations include a relatively small dataset, which may not be broadly generalizable, and the cross-sectional nature of the data, which prevents the establishment of causal relationships. Future research should aim to use larger, more diverse datasets and employ longitudinal data to track changes over time. Exploring more advanced clustering techniques could also provide deeper, more granular insights into the intricate relationships between sleep, health, and overall well-being.

6. Declarations

6.1. Author Contributions

Conceptualization: A.D.S., M.A.W.P.; Methodology: A.D.S., M.A.W.P.; Software: A.D.S.; Validation: M.A.W.P.; Formal Analysis: A.D.S.; Investigation: A.D.S.; Resources: M.A.W.P.; Data Curation: A.D.S.; Writing – Original Draft Preparation: A.D.S.; Writing – Review and Editing: M.A.W.P.; Visualization: A.D.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Lallukka, B. Sivertsen, E. Kronholm, Y. S. Bin, S. Øverland, and N. Glozier, "Association of Sleep Duration and Sleep Quality With the Physical, Social, and Emotional Functioning Among Australian Adults," *Sleep Health*, 2018, doi: 10.1016/j.sleh.2017.11.006.

[2] A. S. Purwanti, A. C. Nabilla, and R. Maulina, "The Effect of Application of Sleep Training Methods With Positive Routines on the Quality of Sleep of Infants Aged 3-4 Months at the Village Health Centre Mulyogagung - Dau Malang District," *Siklus Journal Research Midwifery Politeknik Tegal*, 2024, doi: 10.30591/siklus.v13i01.5689.

[3] A. J. Scott, T. L. Webb, M. M. James, G. Rowse, and S. Weich, "Improving Sleep Quality Leads to Better Mental Health: A Meta-Analysis of Randomised Controlled Trials," *Sleep Medicine Reviews*, 2021, doi: 10.1016/j.smrv.2021.101556.

[4] M. B. Becerra, R. J. Gumasana, J. A. Mitchell, J. B. Truong, and B. J. Becerra, "COVID-19 Pandemic-Related Sleep and Mental Health Disparities Among Students at a Hispanic and Minority-Serving Institution," *International Journal of Environmental Research and Public Health*, 2022, doi: 10.3390/ijerph19116900.

[5] L. Freeman and N. C. Gottfredson, "Using Ecological Momentary Assessment to Assess the Temporal Relationship Between Sleep Quality and Cravings in Individuals Recovering From Substance Use Disorders," *Addictive Behaviors*, 2018, doi: 10.1016/j.addbeh.2017.11.001.

[6] C.-W. Fan, K. Drumheller, I. Chen, and H. Huang, "College Students' Sleep Difficulty During COVID-19 and Correlated Stressors: A Large-Scale Cross-Sessional Survey Study," *Sleep Epidemiology*, 2021, doi: 10.1016/j.sleepe.2021.100004.

[7] C. Du *et al.*, "Health Behaviors of Higher Education Students From 7 Countries: Poorer Sleep Quality During the COVID-19 Pandemic Predicts Higher Dietary Risk," *Clocks & Sleep*, 2021, doi: 10.3390/clockssleep3010002.

[8] L. Pillay *et al.*, "Nowhere to Hide: The Significant Impact of Coronavirus Disease 2019 (COVID-19) Measures on Elite and Semi-Elite South African Athletes," *Journal of Science and Medicine in Sport*, 2020, doi: 10.1016/j.jsams.2020.05.016.

[9] H. M. Milojevich and A. F. Lukowski, "Sleep and Mental Health in Undergraduate Students With Generally Healthy Sleep Habits," *Plos One*, 2016, doi: 10.1371/journal.pone.0156372.

[10] A. Angehrn, M. J. N. Sapach, R. Ricciardelli, R. S. MacPhee, G. S. Andérson, and R. N. Carleton, "Sleep Quality and Mental Disorder Symptoms Among Canadian Public Safety Personnel," *International Journal of Environmental Research and Public Health*, 2020, doi: 10.3390/ijerph17082708.

[11] N. Malik, I. Ashiq, and R. M. Khan, "Sleep Quality and Sleep Hygiene as Predictors of Mental Health Among University Students," *J. Asian Dev. Studies*, 2024, doi: 10.62345/jads.2024.13.1.560.

[12] A. T. Rayward, M. J. Duncan, W. J. Brown, R. C. Plotnikoff, and N. W. Burton, "A Cross-Sectional Cluster Analysis of the Combined Association of Physical Activity and Sleep With Sociodemographic and Health Characteristics in Mid-Aged and Older Adults," *Maturitas*, 2017, doi: 10.1016/j.maturitas.2017.05.013.

[13] S. A. Creasy *et al.*, "Higher Amounts of Sedentary Time Are Associated With Short Sleep Duration and Poor Sleep Quality in Postmenopausal Women," *Sleep*, 2019, doi: 10.1093/sleep/zsz093.

[14] D. Meyer *et al.*, "Health Behaviors and Health Status in Couples Experiencing Pandemic Stress," *The Family Journal*, 2024, doi: 10.1177/10664807241226697.

[15] Y. Zhu, J. Huang, and M. Yang, "Association Between Chronotype and Sleep Quality Among Chinese College Students: The Role of Bedtime Procrastination and Sleep Hygiene Awareness," *International Journal of Environmental Research and Public Health*, 2022, doi: 10.3390/ijerph20010197.

[16] S. Hershner and L. M. O'Brien, "The Impact of a Randomized Sleep Education Intervention for College Students," *Journal of Clinical Sleep Medicine*, 2018, doi: 10.5664/jcsm.6974.

[17] J. A. Nutakor *et al.*, "The Relationship Between Social Capital and Sleep Duration Among Older Adults in Ghana: A Cross-Sectional Study," *International Journal of Public Health*, 2023, doi: 10.3389/ijph.2023.1605876.

[18] X. Chen *et al.*, "Relationship Between Sleep Duration and Sociodemographic Characteristics, Mental Health and Chronic Diseases in Individuals Aged From 18 to 85 Years Old in Guangdong Province in China: A Population-Based Cross-Sectional Study," 2020, doi: 10.21203/rs.3.rs-38654/v1.

[19] C. Del-Valle-Soto, R. A. Briseño, L. J. Valdivia, R. Velázquez, and J. A. Nolazco-Flores, "Non-Invasive Monitoring of Vital Signs for the Elderly Using Low-Cost Wireless Sensor Networks: Exploring the Impact on Sleep and Home Security," *Future Internet*, 2023, doi: 10.3390/fi15090287.

[20] M. Hansen, K. R. Simon, J. Strack, X. He, K. G. Noble, and E. C. Merz, "Socioeconomic Disparities in Sleep Duration Are Associated With Cortical Thickness in Children," *Brain and Behavior*, 2022, doi: 10.1002/brb3.2859.

[21] S. Kļaviņa-Makrecka, I. Gobiņa, I. Pudule, B. Velika, D. Grīnberga, and A. Villeruša, "Poor Self-Reported Health in Association With Sleep Duration and Health Complaints Among Adolescents in Latvia," *SHS Web of Conferences*, 2024, doi: 10.1051/shsconf/202418402003.

[22] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," *Ieee Access*, 2020, doi: 10.1109/access.2020.2988796.

[23] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J — Multidisciplinary Scientific Journal*, 2019, doi: 10.3390/j2020016.

[24] B. N. Sari, "Identification of Tuberculosis Patient Characteristics Using K-Means Clustering," *Scientific Journal of Informatics*, 2016, doi: 10.15294/sji.v3i2.7909.

[25] A. Setiyaji and H. D. Purnomo, "Analyzing the Distribution of Health Workers in Semarang City Using K-Means Clustering Method," *Journal of Information Systems and Informatics*, 2024, doi: 10.51519/journalisi.v6i1.663.

[26] Berlilana and A. Mu'amar, "Economic Decentralization through Blockchain Opportunities Challenges and New Business Models," *Journal of Current Research in Blockchain*, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/jcrb.v1i2.14.

[27] A. Sharannavar and N. B. P K, "Performance Analysis of Clustering Algorithms for Dyslexia Detection," *Ecs Transactions*, 2022, doi: 10.1149/10701.10021ecst.

[28] A. S. Paramita, "Improving K-Nn Internet Traffic Classification Using Clustering and Principle Component Analysis," *Bulletin of Electrical Engineering and Informatics*, 2017, doi: 10.11591/eei.v6i2.608.

[29] P. Yıldırım, G. Başol, and A. Y. Karahan, "Pilates-Based Therapeutic Exercise for Pregnancy-Related Low Back and Pelvic Pain: A Prospective, Randomized, Controlled Trial," *Turkish Journal of Physical Medicine and Rehabilitation*, 2022, doi: 10.5606/tftrd.2023.11054.

[30] M. E. Patrick, J. Griffin, E. D. Huntley, and J. L. Maggs, "Energy Drinks and Binge Drinking Predict College Students' Sleep Quantity, Quality, and Tiredness," *Behavioral Sleep Medicine*, 2016, doi: 10.1080/15402002.2016.1173554.

[31] C. Lewis, K. Lewis, N. J. Kitchiner, S. Isaac, I. Jones, and J. I. Bisson, "Sleep Disturbance in Post-Traumatic Stress Disorder (PTSD): A Systematic Review and Meta-Analysis of Actigraphy Studies," *European Journal of Psychotraumatology*, 2020, doi: 10.1080/20008198.2020.1767349.

[32] J. X. Teo *et al.*, "Digital Phenotyping by Consumer Wearables Identifies Sleep-Associated Markers of Cardiovascular Disease Risk and Biological Aging," *Communications Biology*, 2019, doi: 10.1038/s42003-019-0605-1.

[33] R. Alotaibi, M. Alshammari, D. S. Alotaibi, N. Al-Ansary, S. Acharya, and F. Albagmi, "Association Between Sleep Quality and Its Effect on General Health Among First Year University Students in Saudi Arabia," 2022, doi: 10.21203/rs.3.rs-2193770/v1.