Departamento de Ingeniería y Ciencias. 01, NO. 01, November 2025

Delving into Sleep Quality Using PCA

Pablo E. González-Rios[1], Angelo D. Farfán-Torres[2], Juan P. Enríquez-Ruíz-Velasco[3], Jun K. Lee-Yang[4], Eduardo Medina-Meza[5] and Jenaro Alcaraz-Saragosa[6]

[1] *Instituto Tecnológico de Estudios Superiores de Monterrey (ITESM), Departamento de Ingeniería y Ciencias, Jalisco, México*

Publication date: 28/11/2025

Abstract— This study inspects the relationships that certain habits and factors a person has in relation to their quality of sleep. In order to do so, this study employed Principal Component Analysis (PCA) to explore the structure of an aggregated health and lifestyle dataset and to determine which underlying latent factors, or principal components, exhibit the strongest correlation with self-reported Quality of Sleep. This was needed because there are many variables involved in the result of quality of sleep. The goal was to identify the most critical health and behavioral determinants of sleep quality for potential intervention targeting.

Keywords— Principal Component Analysis (PCA), Quality of Sleep, Data Analysis, K-nearest neighbor

I. Introduction

a. Motivation

Even though it is known that sleep quality influences cognitive performance, metabolic health and long term well-being, a good rest is often neglected by most adults. Poor sleep quality is associated with high levels of stress, which contributes to decreased productivity in the workplace and higher risk of contracting illnesses [1]. However, sleep quality is influenced by multiple factors, such as lifestyle habits, work patterns, or mental health, making it challenging to identify primary causes. Given the above, we consider it essential to thoroughly understand which factors are strongly associated with a poor sleep quality [2].

b. Objectives

The **effective** analysis of high-dimensional data remains a central challenge. This exploratory study applies **Principal Component Analysis (PCA)** to understand patterns within datasets with more than two dimensions. PCA allows the reduction of dimensionality, by transforming a large set of correlated variables into orthogonal linear combinations called Principal Components (PCs), making it easier to **understand complex relationships**. Using a sleep and health database as a case study, we explore how PCA can uncover underlying factors that influence self-reported sleep quality. We aim to: **(1)** reduce dimensionality by applying PCA to the selected database, and **(2)** identify which factors most strongly predict sleep quality.

II. Methodology

a. Dataset

This study utilizes the publicly available Sleep Health and Lifestyle Dataset obtained from Kaggle [3]. The dataset comprises 374 observations from adults aged 27 to 59 years. Participants self-reported various health and lifestyle metrics including gender, age, occupation, sleep quality, physical activity level, stress level, BMI category, and blood pressure. Quality of Sleep is the primary variable for this analysis. Table 1 provides a detailed description of all variables.

Table 1: Description of variables

Variable	Type	Range
Gender	Cat.	M, F
Age	Cont.	27-59 yrs
Occupation	Cat.	Various
Quality of Sleep	Ord.	1-10
Physical Activity	Cont.	0-100 min/day
Stress Level	Ord.	1-10
BMI Category	Cat.	Normal/Overweight/Obese
Blood Pressure	Cat.	Systolic/Diastolic
Sleep Duration	Cont.	5.8-8.5 hrs
Daily Steps	Cont.	3000-10000
Heart Rate	Cont.	65-86 bpm
Sleep Disorder	Cat.	None, Insomnia, Sleep Apnea

b. Data Preprocessing

Data pre-processing involved several steps to prepare the dataset for analysis. First, we removed the person ID column as it contained no analytical value.

Variable transformation: Blood pressure, originally stored as a combined categorical variable (e.g., “120/80”), was split into two continuous variables: Systolic Pressure and Diastolic Pressure. This transformation enabled more granular analysis of cardiovascular metrics.

Encoding categorical variables: Categorical variables including Gender, Sleep Disorder, and BMI Category were encoded as integers ranging from 0 to $k-1$, where k represents the number of unique categories.

Clinical grouping: To facilitate interpretation, we recoded several variables into clinically meaningful categories:

- **Stress Level:** Original values (3–4) were recoded as "Low," (5–6) as "Moderate," and (7–8) as "High."
- **Quality of Sleep:** Values (4–6) were recoded as "Fair," (7) as "Good," and (8–9) as "Excellent."
- **Occupation:** Related professions were consolidated into broader categories: Nurses and Doctors became "Healthcare Professional"; Engineers, Software Engineers, and Scientists became "STEM Professional"; and Managers, Sales Representatives, and Salespersons became "Sales Professional." Teachers, Accountants, and Lawyers remained as individual categories.

Following these transformations, all categorical variables were numerically encoded for compatibility with PCA and multiple machine learning algorithms.

To gain initial insights from the relationships among variables, we constructed a correlation matrix of all processed features. This preliminary visualization helped us intuit about which variables may cluster together in the subsequent PCA.

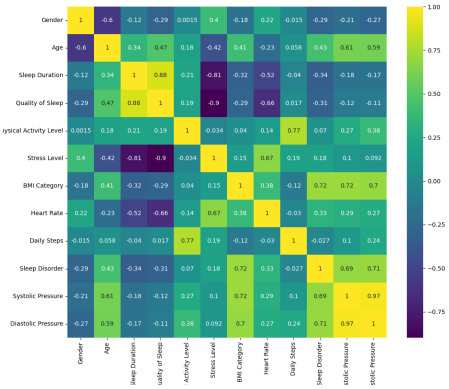


Fig. 1: First correlation matrix

c. Statistical Methods

1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a **dimensionality-reduction technique**, widely used in data analysis and machine learning. Its main purpose is to transform a high-dimensional data set into a lower-dimensional data set, while attempting to preserve the highest amount of variance. This is achieved by constructing new, uncorrelated variables known as **PCs** (Principal Components). These are just linear combinations of the original variables.

The first principal component, **PC1**, captures the maximum amount of variance in the data, while the second principal component, **PC2**, captures the next highest variance, and so forth. The impressive part of PCA is that **ALL** of the principal components are orthogonal to each other. This way, we can create a representation of the data, with the amount of dimensions that we wish for.

By retaining the top n amount of components (those that account for the majority of the data's variability), it effectively reduces redundancy, enhances computational efficiency and facilitates visualization.

In order to proceed with a PCA, the data must first be standardized to have a mean of 0 and standard deviation of 0. Next, a covariance matrix is computed to quantify how the original features vary together. With this matrix, the eigenvalues and eigenvectors are derived. Using the eigenvectors that correspond to the highest eigenvalues, are the principal components. Finally, the data is projected into this reduced subspace [4].

2. K-nearest neighbors algorithm (KNN)

The K-Nearest Neighbors (KNN) algorithm is a machine learning method that is useful for classification and regression. It is described as a lazy-learner: it stores the dataset and only employs it when necessary. When presented with a new data point, it identifies the k nearest points (also referred to as neighbors), and determines the output of this new data point depending on the labels of the k amount of nearest points.

An critical parameter for this algorithm is the value for k. If it is too small, it can result in the prediction to become overly sensitive to outliers or noise. This is known as **overfitting**. On the other hand, if the value for k is too large, it may oversmooth predictions by diluting local patterns with distant points. This is known as **underfitting**.

The way the distance between points is calculated is through **Euclidean distance**, i.e: $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ in a 2-dimensional plane. Now, a new data point can be entered into the KNN algorithm, and it will predict in which "group" of data the new data point belongs to.

However, the KNN algorithm faces a few challenges. Its computational cost scales poorly with large datasets, since a point's distance requires comparing the test point to all training instances. Additionally, in high-dimensional spaces, distance metrics lose discriminative power. And lastly, without proper tuning, an improper k value can result in overfitting, underfitting or no relevance in a dataset with much noise [5].

III. Results

To perform the **PCA**, we used the PCA library from sklearn.decomposition in python. Using this PCA, given the cumulative variance of 0.96 (representation of 96% of the data), we get the following variables:

1. Gender
2. Age
3. Sleep Duration
4. Stress Level
5. BMI Category
6. Heart Rate
7. Physical Activity Level

Using this PCA , we were able to retain 64.94% of the original feature's variance (as mentioned above). To visualize the results, PC1 and PC2 were graphed on a two-dimensional space. They are the first two components of the transformed data, and are a combination of all the variables:

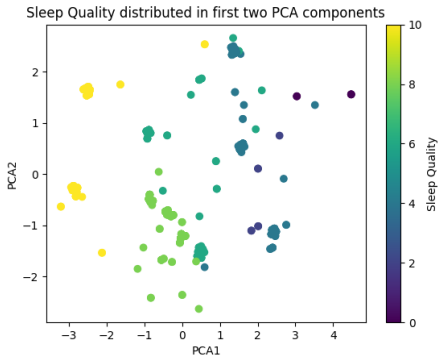


Fig. 2: PCA graph with the most important variables

In order to better visualize our data, we decided to also use the **T-distributed Stochastic Neighbor Embedding** with the relevant variables. This algorithm allows us to reduce many dimensional variables it into a lower dimensional space for a cleaner visualization.

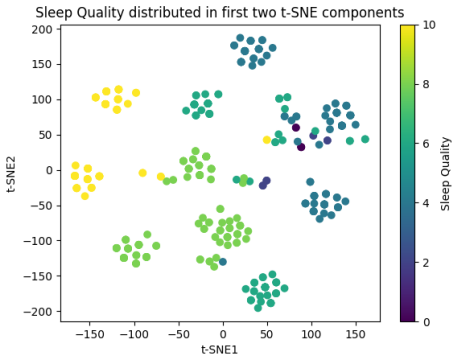
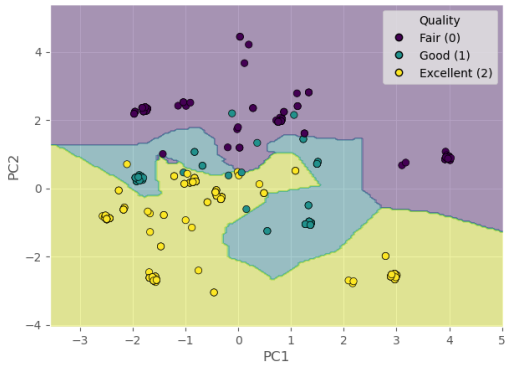


Fig. 3: t-SNE analysis with the relevant variables

After using PCA and t-SNE to better visualize the data, we evaluated multiple machine learning algorithms for sleep quality classification. K-Nearest Neighbors (KNN) with k=3 yielded the best performance when applied to the first two principal components. The K value (k = 3) for the PCA -based KNN model was selected manually through a trial-and-error process, adjusting k until acceptable separation and predictive behavior were observed.

PC1 (Cardiovascular & Metabolic Health) explained 33% of the total variance and was primarily loaded on age, BMI category, sleep disorder, and both systolic and diastolic blood pressure. PC2 (Stress & Mental Health)accounted for an additional 23% of the variance, with dominant contributions from heart rate, stress level, sleep duration, age, and gender. Together, these two components captured 56% of the total variability in the dataset. Note that we considered variables with absolute PCA loadings greater than 0.4 as influential.

Despite retaining around half of the original variance, this reduced feature space proved to be highly effective for classification. Using a train-test split of 299:75 observations (80:20 ratio), the KNN model achieved a classification accuracy of 0.96 (96%). The following figure illustrates the decision boundaries and distribution of sleep quality categories in this two-dimensional PCA space, demonstrating clear separability between quality levels.



ber 2025

Fig. 4: KNN model using the PCA components

Although the model may exhibit overfitting due to the low k value and the projection onto principal components, we argue that the decision remains valid. The classification behavior observed is stable and consistent with the interpretation of latent factors, supporting that this empirical choice still provides a strong and meaningful assertion, as can be seen in the confusion matrix.

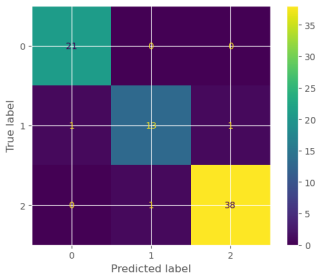


Fig. 5: Confusion matrix for our Knn algorithm

Therefore, we can say that individuals with poorer sleep quality tend to appear at high PC2 (driven by stress and elevated heart rate) and at extreme PC1 values (linked to higher blood pressure, sleep disorders, and overweight/obesity BMI categories). This aligns with our goal of identifying the determinants of poor sleep quality, showing that stress and cardiovascular-metabolic burden are predictive keys for bad rest.

IV. Discussion and Conclusions

Some formal studies have explored the link between sleep quality and physical and mental factors, reaching conclusions that align closely with our PCA-based Knn prediction model. For example, the study by Sajjadieh et al. (2020) [6] found that poor sleep quality correlated with higher heart rate, elevated blood pressure, and increased emotional stress in adults. Also, the study by Kanki et al. (2024) [7] was able to conclude that physical features like blood pressure, heart rate, and BMI significantly affect sleep quality.

In conclusion, our study was able to apply PCA to identify the key components of sleep quality in a health and lifestyle dataset provided by a public database. Our analysis revealed that stress and cardiovascular healthcare the keys of poor sleep quality, with our KNN classification model achieving 96% accuracy. Future work should explore additional models and studies to validate our results on mostly on larger and more controled datasets.

ACKNOWLEDGEMENTS

We acknowledge Laksika Tharmalingam for making the Sleep Health and Lifestyle Dataset publicly available on Kaggle.

This analysis was conducted using Python and open source libraries such as scikit-learn, pandas, matplotlib, and numpy.

This work was completed as part of Analysis of Mathematical Methods for Physics at Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Guadalajara.

REFERENCES

[1] M. A. G. W. P. D. A. J. P. Alberto R. Ramos, MD, "Sleep deprivation, sleep disorders, and chronic disease," 2023. Last accessed November 27 2025.

[2] N/A, "Sleep deprivation impacts," 2022. Last accessed November 25 2025.

[3] L. Tharmalingam, "Sleep health and lifestyle dataset," 2023. Last accessed November 28 2025.

[4] GeeksforGeeks, "Principal component analysis (pca)," 2025. Last accessed November 28 2025.

[5] GeeksforGeeks, "K-nearest neighbor(knn) algorithm," 2025. Last accessed November 28 2025.

[6] A. Sajjadieh, F. Ghassemi, A. Rezaei, M. Dehghani, and S. Azizi, "The relationship between sleep quality with blood pressure, heart rate and emotional intelligence in adults," 2020. Last accessed November 28, 2025.

[7] M. Kanki, Y. Onishi, N. Ito, H. Kose, and T. Araki, "Identification of sleep quality determinants using wearable data and machine learning," 2024. Last accessed November 28, 2025.

A. APPENDICES

The code and analysis for this project are available in our repository: *GitHub*.
`https://github.com/Juamops/Reto_F1009`