Research for the Effects of Sleep Quality Based on Multiple Linear Regression and Random Forest Model

Zhuoyu Long

Beijing 101 Middle School, Beijing, 100091, China

writing@pathacademics.org

Abstract. Sleep efficiency is a key factor in human health. However, existing research tends to focus on isolated variables, ignoring the intricate relationship between sleep patterns and daily habits. In this paper, the Sleep Efficiency Dataset of the Kaggle platform was selected to analyze the effects of sleep structure and lifestyle factors on sleep efficiency. Through multiple linear regression models, the relationship between sleep duration, deep sleep percentage, light sleep percentage, rapid eye movement (REM) sleep percentage, and number of wake-ups was explored. The results showed that deep sleep and REM sleep percentage had a significant positive effect on sleep efficiency, while the number of wake-ups had a negative effect. The random forest model further analyzed the effect of lifestyle on sleep efficiency, and the results showed that age, alcohol consumption, and smoking status were the main factors affecting sleep efficiency. The results of the study provide data to support the improvement of sleep efficiency.

Keywords: Multiple linear regression; Random forest; Sleep quality.

1. Introduction

Sleep quality is essential for the body's physical and mental health. According to a 2024 global sleep survey by ResMed that included 36,000 participants from 17 countries, about 56% of respondents reported lower sleep quality [1]. Studies have shown that poor sleep efficiency further exacerbates the risk of death [2]. Therefore, it is of great significance to study the factors affecting sleep efficiency to improve sleep quality and ensure physical and mental health.

In an existing study, Kim surveyed 125 Korean long-term care residents and explored the effects of multiple factors on sleep efficiency and sleep quality based on multiple linear regression (MLR) models [3]. The study showed that these factors combined affected the sleep of long-term care residents, providing a basis for the development of a comprehensive intervention program [3]. Similarly, Zhang used a stepwise regression analysis to explore the relationship between sleep quality scores, sleep efficiency, and lifestyle factors, and the results showed that sleep efficiency was closely related to exercise frequency, alcohol consumption, and smoking status [4]. Research has shown that many life factors that may be closely related to sleep have not received sufficient attention in previous studies. As a result, sleep efficiency can be affected by several complex factors. Ma's study noted that there may be a nonlinear relationship between many factors that affect sleep [5]. Combined with more powerful models, such as random forests (RF), potential nonlinear relationships can be effectively captured. In addition, although the above findings can help improve sleep quality, how to effectively implement these interventions is still a question worth exploring.

Based on this, this paper analyzes the impact of lifestyle on sleep efficiency by RF model and analyzes the impact of sleep structure on sleep efficiency by using the MLR model. The aim of the study was to explore the key influencing factors of sleep efficiency through these two methods and make effective recommendations for improving sleep quality.

2. Method

2.1. Dataset

This paper selects a set of Sleep Efficiency Datasets from Kaggle. The dataset provides detailed sleep pattern information for 432 subjects, including sleep efficiency, the proportion of each sleep

stage (e.g., rapid eye movement (REM) sleep, deep sleep, and light sleep, etc.), and various lifestyle factors that affect sleep (e.g., caffeine, alcohol, smoking, and exercise, etc.). The sample size of the dataset is sufficient, and the comprehensive variables provide strong data support for the analysis of this study. At the same time, Kaggle is a well-known data science platform, and its datasets are often rigorously vetted and validated, ensuring the accuracy and completeness of the data. The dataset has a clear structure and contains a variety of variables that can be quantified and analyzed, which has high academic value.

2.2. Experimental Design

2.2.1 Data processing

First, this paper converts categorical variables in the data, such as gender and smoking status, into numeric types so that the model can handle them. Next, the missing values in the data were processed and the integrity of the data was guaranteed by filling them with the mean of each column. At the same time, check and remove outliers from the data to ensure that there are no infinity or extreme values in the data. These values can affect the stability and prediction of the model. Finally, the data is processed using standardized techniques to ensure that the scales between features are consistent, to avoid the excessive impact of some features on model training

2.2.2 MLR

In this paper, MLR models were selected to analyze the effects of sleep structure on sleep efficiency. MLR models can reveal the linear relationship between the dependent variable and multiple independent variables, and quantify the influence of independent variables on the dependent variable through regression coefficients. The model is relatively simple to calculate and is suitable for dealing with a large number of independent variables. The specific formula is as follows:

$$\begin{aligned}\text{Sleep efficiency} =& \beta_0 + \beta_1(\text{sleep duration}) + \beta_2(\text{deep sleep percentage}) + \beta_3(\text{light sleep percentage}) \\ &+ \beta_4(\text{REM sleep percentage}) + \beta_5(\text{wake-up times}) + \varepsilon\end{aligned} \tag{1}$$

β_0 is the intercept term, which represents the expected value of sleep efficiency when all independent variables are zero. β_1, β_2, β_3, β_4, β_5 are regression coefficients, which represent the degree and direction of the influence of each variable on sleep efficiency. ε is the error term, which represents the random factors that the regression model cannot explain.

2.2.3 RF

The RF regression model predicts through the integration of multiple decision trees, has strong nonlinear modeling capabilities, the ability to process high-dimensional data, and can evaluate the importance of characteristic variables. In this study, a RF model was used to analyze the effect of lifestyle on sleep efficiency. In this study, firstly, the target variable sleep efficiency was separated from other characteristic variables, and the dataset was divided into a training set and a test set, with 80% of the data used for training and 20% for testing. After training, the model evaluates the relative importance of individual features.

To verify the validity of the model, the mean square error (MSE) and mean absolute error (MAE) are used to measure the prediction error of the model, and the R^2 value is used to evaluate the degree to which the model fits the test data. To evaluate the stability of the model more comprehensively, a cross-validation method is used, which can effectively avoid the problem of model overfitting by dividing the data into multiple subsets and training and evaluating the model multiple times.

3. Result

Table 1 shows the results of multiple regression models to analyze the influencing factors of sleep efficiency. The coefficient of determination (R-squared) of the model is 0.772, indicating that the model has strong explanatory power. The adjusted R-squared was 0.770, the F-statistic was 362.0,

and its significance level was less than 0.005, indicating that the overall regression model was statistically significant. In addition, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were -1184 and -1164, respectively, indicating that the model had a better fit when compared with other models.

Table 1. MLR results

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0002	0.0000	27.4547	0.0000	0.0002	0.0003
Sleep duration	0.0006	0.0036	0.1548	0.8771	-0.0066	0.0077
Deep sleep percentage	0.0096	0.0003	27.9626	0.0000	0.0089	0.0103
REM sleep percentage	0.0112	0.0007	15.2668	0.0000	0.0097	0.0126
Light sleep percentage	0.0033	0.0004	8.4217	0.0000	0.0025	0.0040
Awakenings	-0.0336	0.0024	-13.7821	0.0000	-0.0384	-0.0288

The results in Table 1 show that the percentage of deep sleep and the percentage of REM sleep have a significant positive effect on sleep efficiency, while the percentage of light sleep also has a positive effect on sleep efficiency, but to a lesser extent. In addition, the number of wake-ups had a significant negative effect on sleep efficiency. In contrast, the relationship between sleep duration and sleep efficiency was not significant. Overall, the model had a high degree of fit and could better explain the changes in sleep efficiency. Figure 1 shows the results of the RF analysis, which shows the importance of different factors to sleep efficiency. According to the results of RF analysis, age, alcohol consumption, and smoking status were the main factors affecting sleep efficiency, while sleep duration, caffeine consumption, and gender had a small effect.

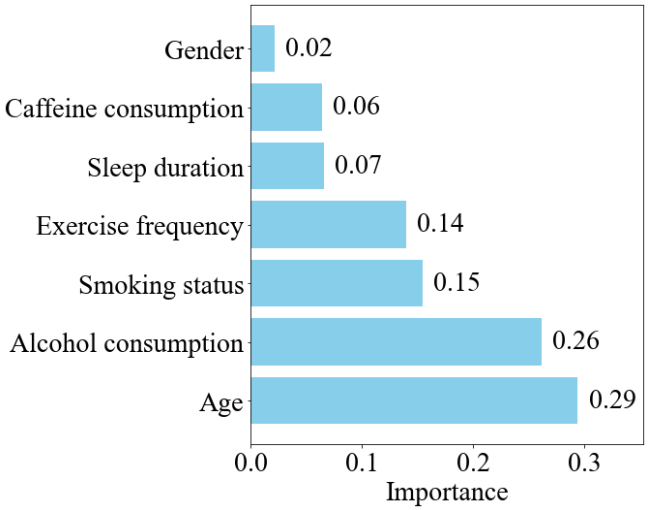

Figure 1. RF analysis results (Photo/Picture credit: Original).

Table 2 shows the results of the evaluation of the performance of the RF model. The results show that the prediction error of the model is small, but the overall fit is average, and the model can explain about 39.2% of the variability.

Table 2. Evaluation results of the performance of the RF model

MSE	MAE	R^2 Score
0.0115	0.0894	0.3917

4. Discussion

Based on the results of this study, it is recommended to increase the ratio of deep sleep and REM sleep, and reduce the number of awakenings to improve sleep efficiency. Studies have shown that the percentage of deep sleep and the percentage of REM sleep have a significant positive effect on sleep efficiency. Deep sleep and REM sleep can be promoted through relaxation techniques, regular sleep

schedules, and avoiding excessive stress and anxiety. At the same time, the number of wake-ups had a significant negative effect on sleep efficiency. To reduce the number of awakenings, it is recommended to avoid irritants (such as caffeine and alcohol) before bedtime and create a quiet, comfortable sleeping environment. In addition, maintaining a regular sleep schedule and improving the comfort of your sleeping environment, such as the right mattress and temperature, can also help reduce nocturnal awakenings [6].

RF model analysis showed that lifestyle factors such as alcohol consumption, smoking, and age had a greater impact on sleep efficiency. Reducing alcohol and smoking and maintaining a moderate amount of physical activity, especially during the day, can help improve sleep quality. However, there are several limitations in the process of conducting experiments and analyses. Although the RF model has the advantage of prediction accuracy, its black-box nature makes the model less explanatory. Although this paper evaluates the importance of features to understand which factors have a greater impact on sleep efficiency, there is still a lack of sufficient explainability for specific variable interactions and mechanism analysis. Future research may consider combining more interpretable models or using feature importance analysis in combination with other methods such as SHAP values to enhance the transparency of the model. Wilson noted that SHAP provides a more accurate and nuanced explanation when considering the interplay between model complexity and features [7].

The dataset used in this paper is mainly from the Kaggle platform, and although the dataset is relatively large and from reliable sources, it is not fully representative of all populations, especially in the context of different cultures, geographical locations, or age groups, and may be biased. Therefore, future studies can further expand the sample scope to include more extensive and diverse population data, to enhance the universality of the study.

5. Conclusion

In this study, MLR and RF models were used to analyze the effects of sleep structure and lifestyle on sleep efficiency, respectively. Based on the analysis results of MLR model, the percentage of deep sleep and REM sleep have a significant positive effect on sleep efficiency, indicating that increasing the proportion of deep sleep and REM sleep can effectively improve sleep efficiency. Light sleep percentage and sleep duration have a small effect on sleep efficiency, although they have a certain effect. The number of wake-ups had a significant negative effect on sleep efficiency, indicating that frequent awakenings significantly reduced sleep efficiency. The coefficient of determination (R^2) of the multiple regression model was 0.772, indicating that the model could explain the changes in sleep efficiency well. Secondly, the RF regression model revealed the effect of lifestyle factors on sleep efficiency. In the feature importance analysis, alcohol consumption, smoking status, and age were identified as the most important factors, indicating that these factors had a greater impact on sleep quality. Comparatively, sleep duration, caffeine consumption, and gender had less effect. The MSE, and MAE of the model were 0.0115, the MAE was 0.0894, and the R^2 was 0.3917, indicating that although the model had strong predictive power, its fit was low, reflecting the complex nonlinear relationship between lifestyle and sleep efficiency.

Taken together, an increase in deep sleep and REM sleep, a decrease in the number of awakenings, and a healthy lifestyle (especially a reduction in alcohol and smoking) are key factors in improving sleep efficiency. However, the black-box nature of RF models limits the in-depth understanding of variable interactions, so future research can be combined with other interpretable methods, such as SHAP value analysis, to further explore the complex relationships between different factors. In addition, the universality and accuracy of the study can be improved in the future by expanding the sample range, especially for groups with different cultural and geographical backgrounds.

References

[1] Ford S J, dos Santos R, dos Santos R. Empowering Female High School Students for STEM Futures: Career Exploration and Leadership Development at Scientella. Education Sciences, 2024, 14(9): 955.

[2] Yuan H, Plekhanova T, Walmsley R, et al. Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. NPJ digital medicine, 2024, 7(1): 86.

[3] Kim D E, Yoon J Y. Factors that influence sleep among residents in long-term care facilities. International Journal of Environmental Research and Public Health, 2020, 17(6): 1889.

[4] Zhang Y. The Impact of Lifestyle Factors on Sleep Efficiency and Sleep Quality. Highlights in Science, Engineering and Technology, 2023, 54: 351-356.

[5] Ma Z, Lin Z. The impact of exposure to memorial reports on the 5.12 Wenchuan earthquake on sleep quality among adult survivors ten years after the disaster: Evidence for nonlinear associations. Comprehensive psychiatry, 2020, 97: 152150.

[6] Caddick Z A, Gregory K, Arsintescu L, et al. A review of the environmental parameters necessary for an optimal sleep environment. Building and environment, 2018, 132: 11-20.

[7] Marcílio W E, Eler D M. From explanations to feature selection: assessing SHAP values as feature selection mechanism. 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI). Ieee, 2020: 340-347.