Determinants of University Students’ Sleep Quality

Xixiang Gaoqiao[a]
Department of Mathematics, Durham University, Durham, DH1 3LE, U.K.

Keywords: Sleep Quality, Behavioural Factors, Predictive Modelling, Linear Regression, Random Forest.

Abstract: University students often experience poor sleep quality due to various academic and lifestyle pressures. This study analyzes a dataset of 500 university students to identify key determinants of sleep quality. This paper examines demographic factors (age, gender, year of study), daily habits (study hours, screen time, caffeine intake, physical activity), and sleep characteristics (sleep duration, bed/wake times). Multiple linear regression and random forest models were used to explore linear and non-linear relationships. The linear regression models showed very low explanatory power in both the full dataset (n = 500) and a filtered dataset (n = 52), suggesting that none of the measured factors had a significant linear impact. The random forest models performed well on the training set but had poor performance on the test set, indicating possible overfitting. However, both models consistently highlighted sleep duration and study hours as relatively important features. These findings imply that unmeasured factors, such as psychological stress, may play a larger role in students’ sleep quality. In conclusion, while lifestyle and demographic variables alone did not strongly predict sleep quality in this sample, the results highlight the need to consider psychological and environmental factors. This paper discusses the implications, limitations (including the lack of psychological/cultural variables and seasonal effects), and provide suggestions for future research and interventions to improve student sleep health.

1 INTRODUCTION

Sleep is essential for students’ health and academic success. However, many university students do not sleep well due to irregular schedules, heavy study loads and social activities. Surveys have reported that 70.6% of university students get less sleep than the recommended 8 hours per night (Hershner and Chervin, 2014). Some studies even found that up to 31–65% of students suffer from poor sleep quality (Zhang et al., 2022). This lack of sleep can harm their learning and health. For example, insufficient sleep leads to worse academic performance and serious physical health outcomes. It is also associated with mental health conditions like depression and anxiety (Zhang et al., 2022). In short, poor sleep among students is common and it has negative consequences for multiple aspects of their daily lives. Understanding which factors have the greatest influence on sleep quality is important for designing effective interventions and health policies.

The aim of this paper is to investigate the determinants of university students’ sleep quality using a comprehensive dataset. The dataset is fetched from the Kaggle website (Student Sleep patterns) which contains 500 groups of data without any missing data. Each student record contained demographic information (age, gender, and year of study), behavioural factors (average study hours per day, daily screen time, caffeine intake, and weekly physical activity), and sleep-related variables (average sleep duration in hours, a self-reported sleep quality score from 1 to 10, and typical sleep/wake times on weekdays and weekends). In addition, two new variables-Weekday Sleep Duration and Weekend Sleep Duration-were created to better reflect sleep patterns. A filtered dataset (n = 52) was also prepared by removing records with unrealistic or inconsistent sleep schedule data. This paper examines how demographic characteristics and daily habits relate to students’ self-rated sleep quality. Multiple linear regression and random forest models were applied to identify key predictors. The results aim to clarify which factors are most influential for student

[a] https://orcid.org/0009-0001-4307-7716

Gaoqiao, X.
Determinants of University Students’ Sleep Quality.
DOI: 10.5220/0013848000004708
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Conference on Innovations in Applied Mathematics, Physics, and Astronomy (IAMPA 2025)*, pages 631-637
ISBN: 978-989-758-774-0
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

sleep and to provide insights that may guide future interventions. The following sections present a review of relevant literature, the methodology of the analysis, the results obtained, and a discussion of conclusions, limitations, and future outlook.

2 RELATED WORKS

The topic of students' sleep quality has received a lot of interest in recent years. For instance, Lund et al. conducted a large survey of college students in the United States and found that over 60% of students were categorized as poor-quality sleepers according to the Pittsburgh Sleep Quality Index (PSQI). In that study, students commonly had delayed bedtimes on weekends and reported that stress from academics and emotions negatively affected their sleep (Lund et al., 2010; Yang et al., 2012).

Moreover, Becker et al. surveyed over 7,600 students across multiple universities and found that 62% met the criteria for poor sleep quality. Interestingly, they observed some gender differences (with female students reporting slightly higher rates of sleep problems than males) and found that mental health symptoms were strongly associated with sleep issues (Becker et al., 2018). Similarly, a study in Ethiopia by Lemma et al. reported 55.8% of university students as having poor sleep quality (PSQI > 5). That study found that female students and those in higher years of study had higher odds of poor sleep, and importantly, higher perceived stress and symptoms of depression/anxiety were strongly correlated with worse sleep quality (Lemma et al., 2014).

Finally, recent research has started to examine other potential determinants of student sleep quality beyond psychological stress. Schmickler et al. conducted a cross-sectional study with university students in Germany to identify what influences sleep quality. They found nearly half of the students (48.7%) had poor sleep quality (PSQI > 5). Their regression analysis indicated that several factors significantly predicted poor sleep: older students had worse sleep, and students who reported higher stress and exhaustion levels also had lower sleep quality (Schmickler et al., 2023).

Most of these studies used regression analysis or basic statistical tests to find significant factors. However, fewer studies have compared traditional models with machine learning methods to see if prediction can be improved. Overall, previous literature suggests that poor sleep is common among students especially age and gender, mental health and stress are often the main predictors. However, fewer studies have examined whether common daily habits (such as screen time, caffeine use, or exercise) can predict sleep quality in a more detailed model.

3 METHODOLOGIES

3.1 Data Source

A dataset of 500 university students is used. The original dataset remained in the .CVS format. The key variables are collected in Table 1:

Table 1: Descriptions of variables used in the dataset.

Variables	Symbol	Description
Sleep Quality	X1	The outcome variable. This was a self-reported rating of overall sleep quality on a scale from 1 (very poor) to 10 (excellent).
Age	X2	The student's age in years.
Gender	X3	Categorical variable with three groups: Female, Male, and Other.
University Year	X4	The student's current year of university (categorical: '1st Year', '2nd Year', '3rd Year', '4th Year').
Sleep Duration	X5	The average number of hours the student sleeps per night.
Study Hours	X6	The average number of hours per day the student spends on academic work.
Screen Time	X7	The average hours per day the student spends on screens.
Caffeine Intake	X8	The typical number of caffeinated beverages the student consumes in a day.
Physical Activity	X9	The student's level of physical activity, measured by minutes of exercise per week.
Weekday Bedtime	X10	The typical time the student goes to bed on weekdays reported in hour of day.
Weekday Wake-up Time	X11	The typical time the student wakes up on weekdays.
Weekend Bedtime	X12	The typical bedtime on weekends.
Weekend Wake-up Time	X13	The typical wake-up time on weekends.

All time-of-day variables (bedtimes and wake times) were recorded in hours using a 24-hour format with decimals. These times allow people to capture differences in sleep schedule between weekdays and weekends. Before analysis, this paper checked and cleaned the data for any inconsistencies or data loss.

In addition to the original variables, two new variables were created to better capture students' sleep patterns: Weekday Sleep Duration and Weekend Sleep Duration. These were calculated by subtracting bedtime from wake-up time separately for weekdays and weekends. Durations less than zero were adjusted by adding 24 to reflect overnight sleep accurately. To ensure that the calculated sleep durations were meaningful and reflected typical sleep behaviour, specific filtering criteria were applied. Bedtimes were required to fall between 19:00 (7 PM) and 5:00 (5 AM), and wake-up times were required to be between 4:00 (4 AM) and 11:00 (11 AM). If a student's bedtime or wake-up time fell outside these ranges, the corresponding sleep duration was considered invalid and marked as missing (Table 2).

Table 2: Descriptions of new variables.

Variables	Symbol	Description
Weekday Sleep Duration	X14	Hours of sleep on a typical weekday night (calculated as Wake - Bedtime)
Weekend Sleep Duration	X15	Hours of sleep on a typical weekend night (calculated as Wake - Bedtime)

After applying this filtering process, only the records with valid Weekday Sleep Duration and Weekend Sleep Duration remained in the dataset. This resulted in a filtered dataset of 52 students, which was used in a separate analysis to compare with the full dataset of 500 students. The full dataset allowed a broader analysis using only the original Sleep Duration variable, while the filtered dataset provided more precise measurements of sleep patterns, enabling deeper insights into weekday and weekend sleep behaviour (Sohn et al., 2012).

For clarity, all independent variables are labeled as X1 to X15, as shown in Table 1 and in Table 2, which lists their corresponding names and descriptions.

Categorical variables, including Gender (X3) and University Year (X4), were transformed into dummy variables for use in the models. For Gender, three variables were labeled: Male (X3_1), Other (X3_2), and Female (X3_3). Although Female is the reference category in the regression models and does not have a coefficient, it is still assigned a code (X3_3) for clarity and consistency in tables. Similarly, for University Year, four variables were labeled: 1st year (X4_1), 2nd year (X4_2), 3rd year (X4_3), and 4th year (X4_4), where 1st year serves as the reference group but is also coded for consistency.

Two main analytical approaches are employed to investigate the determinants of sleep quality. Data analysis was conducted with the help of SPSSAU, an online application for statistical modeling.

3.2 Linear Regression

First, a multiple linear regression model is built with Sleep Quality as the dependent variable. All the other measured factors (age, gender, university year, sleep duration, study hours, screen time, caffeine intake, physical activity, and the sleep schedule variables) were used as predictors. Before fitting the model, categorical variables (Gender and Year of Study) were encoded into dummy variables. For example, Gender was represented with two dummy indicators (Male and Other, with Female as the reference category), and Year of Study was represented with dummy variables for 2nd, 3rd, and 4th year (with 1st year as the reference).

The linear regression calculates a coefficient for each predictor. Standard errors and p-values are calculated for each coefficient to assess statistical significance. The fit of the model was evaluated using the coefficient of determination (R^2), which indicates the proportion of variance in sleep quality explained by the predictors. The model predicts that if certain factors (e.g., sleep duration or screen time) had a strong linear relationship with sleep quality, their coefficients would be significantly different from zero and the R^2 would be notably above 0. This paper expected strong predictors to have significant coefficients (low p-values) and a high R^2.

To provide a comprehensive analysis, the regression was performed on both the full dataset (500 students) and a filtered dataset (52 students). The filtered dataset included only students whose sleep schedule data met strict validity criteria, allowing the inclusion of two additional variables: Weekday Sleep Duration and Weekend Sleep Duration. This comparison aimed to investigate whether higher-quality data would improve the model's explanatory power.

3.3 Random Forest

Secondly, a random forest regression model is applied. A random forest is an ensemble machine learning method that builds many decision trees and averages their predictions. Unlike linear regression, it can capture complex relationships automatically, including non-linear effects and interactions between factors (Tan and Greenwood, 2021).

The same set of predictors is used for the random forest model. To train and evaluate the model, this paper splits the dataset into a training set and a test set. 80% of the data (400 students) is used for training the model and the remaining 20% (100 students) is used for testing. The model was trained with 100 decision trees (n_estimators = 100) using default parameters for depth and splitting criteria. The model recorded the R² on the test set, as well as the root mean squared error (RMSE) which gives an estimate of the average prediction error in the same units as the sleep quality scale. The feature importance scores from the random forest are also extracted. These scores indicate how much each predictor contributed to reducing error in the model's decision trees, giving a ranking of which factors the model found most useful for predicting sleep quality.

As with the linear regression, the random forest was applied to both the full dataset and the filtered dataset to compare the model's performance across different data quality levels. The filtered dataset allowed the random forest to incorporate Weekday Sleep Duration and Weekend Sleep Duration, providing deeper insight into how sleep timing affects sleep quality.

Comparing both methods allowed people to see whether a more complex model like random forest would improve prediction accuracy or like linear regression.

4 RESULTS AND DISCUSSION

4.1 Linear Regression Results

The multiple linear regression model for the full dataset (n = 500) showed very weak explanatory power (Table 3). The R-squared was 0.011 and the adjusted R-squared was -0.019. The F-test result was F (15, 484) = 0.370, p = 0.986, which indicates that the model was not statistically significant. All predictors had p-values above 0.05. Among them, Caffeine Intake (B = -0.011, p = 0.889) and Weekday Sleep End (B = 0.055, p = 0.633) showed slightly larger coefficients, but the results were still not significant. This suggests that the full dataset did not provide useful predictors of sleep quality.

The model using the filtered dataset (n = 52) had a higher R-squared of 0.331 (Table 4), though the adjusted R-squared was slightly negative (-0.003). The F-test was F (17, 34) = 0.990, p = 0.491, which was also not statistically significant. However, some predictors showed stronger effects. Sleep Duration (B = 0.406, p = 0.204) and Weekday Sleep Duration (B = 0.454, p = 0.204) had the largest positive coefficients. Caffeine Intake (B = -0.601, p = 0.073) showed a negative effect and had the smallest p-value, though it was still above 0.05.

Table 3: Linear Regression (*n* = 500)

Variables	B	P Value
Constant	4.869	0.019*
X4_4	0.115	0.771
X4_3	0.277	0.463
X4_2	0.110	0.773
X3_2	-0.110	0.750
X3_1	-0.479	0.140
X2	0.018	0.758
X5	-0.042	0.649
X6	0.050	0.207
X7	0.057	0.721
X8	-0.011	0.889
X9	-0.001	0.823
X10	-0.008	0.740
X12	-0.001	0.970
X11	0.055	0.633
X13	-0.020	0.869
R-squared	0.011	
Adjust R-squared	-0.019	
F Test	F (15, 484) = 0.370, p = 0.986	

Table 4: Linear Regression (*n* = 52)

Variables	B	P value
Constant	4.222	0.575
X4_4	0.644	0.676
X4_3	0.188	0.890
X4_2	-0.359	0.805
X3_2	0.476	0.701
X3_1	-1.238	0.309
X2	-0.215	0.333
X5	0.406	0.204
X6	-0.206	0.187
X7	0.418	0.464
X8	-0.601	0.073
X9	-0.002	0.914
X10	-0.064	0.639
X12	0.098	0.617
X11	-0.310	0.584
X13	0.607	0.365
X14	0.454	0.204
X15	-0.172	0.704
R-squared	0.331	
Adjust R-squared	-0.003	
F Test	F (17, 34) = 0.990	

When comparing the two models, the filtered dataset provided clearer patterns and larger coefficients for key variables like Sleep Duration and Caffeine Intake. Although neither model reached statistical significance, the filtered dataset showed improved model fit and stronger trends. This suggests that cleaning the data and focusing on valid observations can help improve the performance of regression models, even if the results are still limited by sample size.

4.2 Random Forest Results

The random forest model with the full dataset (n = 500) also identified Study Hours (weight = 0.128), Weekday Sleep Start (weight = 0.126), and Sleep Duration (weight = 0.097) as the top contributors. Other variables such as Physical Activity and Weekend Sleep Start had moderate importance, but most of the demographic variables, like University Year and Gender, had very low weights, suggesting minimal influence on sleep quality (Table 5 and 6).

Table 5: Feature Weight from Random Forest Model

Variables	Weight
X4_4	0.011
X4_3	0.010
X4_2	0.010
X4_1	0.010
X3_2	0.008
X3_1	0.011
X3_3	0.013
X2	0.046
X5	0.097
X6	0.128
X7	0.074
X8	0.045
X9	0.092
X10	0.126
X12	0.110
X11	0.103
X13	0.107

Table 6: Model Evaluation (n=500)

Indicator	Training Set	Test Set
R-squared	0.849	-0.058
RMSE	1.156	3.007

In addition to analyzing feature importance, the performance of the random forest models was assessed using R-squared and Root Mean Squared Error (RMSE). For the full dataset (n = 500), the model had an R-squared of 0.849 on the training set but dropped to -0.058 on the test set, indicating overfitting and poor generalization. The RMSE values were 1.156 for the training set and 3.007 for the test set, further confirming that the model did not perform well on unseen data.

Table 7: Feature Weight from Random Forest Model (n=52)

Variables	Weight
X4_4	0.003
X4_3	0.015
X4_2	0.015
X4_1	0.004
X3_2	0.023
X3_1	0.014
X3_3	0.004
X2	0.027
X5	0.080
X6	0.134
X7	0.059
X8	0.109
X9	0.042
X10	0.069
X12	0.050
X11	0.043
X13	0.100
X14	0.138
X15	0.074

Table 8: Model Evaluation (n=52)

Indicator	Training Set	Test Set
R-squared	0.832	0.054
RMSE	1.273	2.323

The random forest model using the filtered dataset (n = 52) revealed that the most important predictors of sleep quality were Study Hours (weight = 0.134), Weekday Sleep Duration (weight = 0.138), Weekend Sleep End (weight = 0.100), Sleep Duration (weight = 0.080), and Weekday Sleep Start (weight = 0.069). These variables contributed the most to the model's prediction accuracy. Notably, Weekday Sleep Duration and Study Hours showed the highest weights, indicating that both the length of sleep and academic workload might have a relatively larger influence on students' sleep quality in the filtered dataset (Table 7 and 8).

For the filtered dataset (n = 52), the random forest model achieved an R-squared of 0.832 on the training set and 0.054 on the test set. The RMSE values were 1.273 (training) and 2.323 (test). While the test set R-squared was still low, it was slightly better than the result from the full dataset, suggesting a marginal improvement in predictive stability when using the filtered data.

Comparing the two models, both results highlight Study Hours and Sleep Duration as consistently important factors. However, the filtered dataset emphasized Weekday Sleep Duration and Weekend Sleep End more strongly, while the full dataset placed greater weight on Weekday Sleep Start. This suggests that when outliers and unrealistic data were removed, the model focused more on total sleep duration as a key predictor, while the full dataset's model was slightly more sensitive to the timing of sleep.

4.3 Comparison with Previous Studies

Some researchers have also analyzed the same dataset. For example, Tapendu used both multiple linear regression and random forest models. In their work, they added several new variables, including Average Sleep Duration and Sleep Onset Time Difference, which aimed to describe students' sleep habits in more detail. Their linear regression model achieved a relatively high R-squared, and their random forest results showed that Sleep Duration, Average Sleep Duration, and Weekday Sleep End were the most important factors influencing sleep quality. These results are generally consistent with the trend found in this study, where sleep duration-related variables also appeared to be key predictors. However, there are clear differences between their approach and mine. This paper focused more on improving data quality through strict filtering. The filtered dataset contained only records with valid sleep schedule data, which made it possible to include two new variables: Weekday Sleep Duration and Weekend Sleep Duration. Although my model performance was lower, the focus was on ensuring that only reliable data were used. These differences in variable selection and data processing may explain why the results are not exactly the same (Tapendu, 2024). Overall, both studies highlight that sleep patterns play an important role in determining sleep quality, but different methods can lead to different outcomes.

5 CONCLUSION

This study explored how student lifestyle factors affect sleep quality using both linear regression and random forest models. The results showed that neither model could accurately predict sleep quality, with low R^2 scores on the test set for both models. Although the random forest model performed well on the training data, its performance dropped significantly on the test set, suggesting overfitting. Still, both methods consistently highlighted sleep duration as one of the most important factors influencing sleep quality.

There are also some limitations. The dataset did not include psychological factors such as stress, anxiety or depression, which may strongly influence sleep, especially for international students who are always influenced by culture shock. Environmental factors like sunlight duration and seasonal changes were also not considered.

Moreover, international students often face extra challenges that can affect their sleep quality. In future research, more detailed data should be collected especially data about mental health, social factors. With better and comprehensive data, models may give more useful and accurate predictions of student sleep quality.

REFERENCES

Becker, S. P., Jarrett, M. A., Luebbe, A. M., Garner, A. A., Burns, G. L., Kofler, M. J. 2018. Sleep in a large, multi-university sample of college students: sleep problem prevalence, sex differences, and mental health correlates. *Sleep health*, 4(2), 174-181.

Hershner, S. D., Chervin, R. D. 2014. Causes and consequences of sleepiness among college students. *Nature and science of sleep*, 6, 73-84.

Lemma, S., Berhane, Y., Worku, A., Gelaye, B., Williams, M. A. 2014. Good quality sleep is associated with better academic performance among university students in Ethiopia. *Sleep & breathing = Schlaf & Atmung*, 18(2), 257-263.

Lund, H. G., Reider, B. D., Whiting, A. B., Prichard, J. R. 2010. Sleep patterns and predictors of disturbed sleep in a large population of college students. *The Journal of adolescent health: official publication of the Society for Adolescent Medicine*, 46(2), 124-132.

Schmickler, J. M., Blaschke, S., Robbins, R., Mess, F. 2023. Determinants of Sleep Quality: A Cross-Sectional Study in University Students. *International journal of environmental research and public health*, 20(3).

Sohn, S. I., Kim, D. H., Lee, M. Y. Cho, Y. W. 2012. The reliability and validity of the korean version of the pittsburgh sleep quality index. *Sleep & Breathing*, 16(3), 803-812.

Tan, C. P. Y., Greenwood, K. M. 2021. Stress, Sleep and Performance in International and Domestic University Students. *Journal of International Students*, 12(1), 81-100.

Tapendu. 2024. Student Sleep Pattern Analysis. *Kaggle.* https://www.kaggle.com/code/iamtapendu/student-sleep-pattern-analysis#Multiple-Regression-Analysis

Yang, P. Y., et al. 2012. Exercise training improves sleep quality in middle-aged and older adults with sleep problems: a systematic review. *Journal of Physiotherapy*, 58(3), 157-163.

Zhang, L., Zheng, H., Yi, M., Zhang, Y., Cai, G., Li, C., Zhao, L. 2022. Prediction of sleep quality among university students after analyzing lifestyles, sports habits, and mental health. *Frontiers in psychiatry*, 13, 927619.