DreamHarmony: Unveiling Sleep Patterns through Lifestyle Modeling and Predictive Analytics

Hoang Q. Nguyen*‡, Md Mahmudul Hassan*‡, Solomon Rukundo†‡
* Korea Institute of Science and Technology
† Korea Research Institute of Standards and Science
‡ Korea National University of Science and Technology (UST)

***Abstract*—This study investigates how daily habits like exercise and technology use are related to sleep quality. After careful data preparation, we used statistical methods, including Pearson [1] and Spearman correlations [2], to identify which lifestyle factors are most connected to good or poor sleep, in terms of duration and quality. We then built various mathematical models, such as linear regression and decision trees, to predict sleep quality based on these factors. Our results show that some daily activities have a strong link to how well people sleep, and our models can explain a significant part of this link. These findings could help guide recommendations for improving sleep based on personal habits.**

***Index Terms*—big data, sleep, health, lifestyle, machine learning, support vector machine**

I. Introduction

The interplay between lifestyle and sleep quality is an increasingly pertinent subject in the realm of public health. To explore this intricate relationship, we embarked on a comprehensive data collection campaign, developing a structured survey that captures a wide array of lifestyle variables alongside sleep metrics. The survey's design allowed for the nuanced capture of daily routines, technological interactions, and exercise habits.

Post-collection, data processing commenced with rigorous cleaning techniques to ensure integrity and analytical viability. Our initial exploratory analysis hinged on the construction of a correlation matrix heatmap, which served as a visual aid to discern potential relationships between variables. This graphical representation provided the preliminary clues necessary for subsequent hypothesis formulation.

Armed with hypotheses, we delved into a variety of statistical tests. Spearman's rank correlation offered insights into monotonic relationships, while Pearson's correlation assessed linear dependencies. For categorical data, the Chi-squared test [3] evaluated associations between discrete variables. These statistical methodologies were pivotal in distilling the factors most relevant to sleep quality.

The culmination of our hypothesis testing informed the selection of variables for model fitting. We implemented a variety of machine learning models, including, but not limited to, regression analyses, decision trees, and support vector machines. Each model was rigorously evaluated to determine its predictive power and relevance to our study.

Our contributions through this research are multifaceted. Initially, we developed a foundational survey engineered to assess sleep and lifestyle variables. Subsequently, we discerned pivotal factors that influence sleep quality by leveraging statistical analysis and hypothesis testing. Furthermore, we harnessed machine learning algorithms to construct models based on these factors. Our study establishes a framework for future exploration, highlighting the necessity for continuous data accumulation to refine and augment model efficacy. We invite researchers and the public alike to contribute to our ongoing work (available on GitHub[1]) through feedback and new applications of our findings.

II. Related Work

A large Japanese cross-sectional study (n = 30,000) found that 28% slept less than 6 hours per night and 65% slept less than 7 hours per night despite 80% claiming adequate sleep [4]. Logistic regression identified short sleep duration (<6 hours) as associated with being female, younger age, urban living, unemployment, poor health, lack of exercise, and irregular eating [4]. A review of Asian literature linked insufficient sleep to high stress, high BMI/obesity, and depression [5]. A study of Lithuanian university students (n = 405) found poor sleep quality, especially among medical students, to be associated with high academic demands, anxiety, limited leisure time, and academic dissatisfaction [6]. These findings suggest that interventions targeting stress, leisure time, exercise, urban planning, mental health, and other lifestyle factors could potentially improve sleep [4], [6]. However, existing research lacks exploration of newer lifestyle factors, such as the impact of technology and social media on sleep habits. Future research is needed to investigate how emerging habits like nighttime electronic device use, pre-sleep social media

[1] https://github.com/johnkimtech/sleep

engagement, and internet overuse might contribute to sleep disturbances.

III. Data Acquition & Preprocessing

The DreamHarmony Sleep Survey was an integral part of our study, designed to explore the multifaceted relationship between lifestyle habits and sleep quality. To capture a diverse set of responses, the survey was made available in multiple languages including Korean, English, Vietnamese, and Bengali.

A. Data collection through online survey

The DreamHarmony Sleep Survey was meticulously designed to delve into the myriad factors affecting sleep quality and patterns. This comprehensive survey aimed to unravel the complex interplay between lifestyle habits and sleep, potentially leading to enhancements in sleep quality and overall life satisfaction.

Survey Structure and Distribution: The survey was structured into distinct sections: demographics, lifestyle, and sleep habits. It was distributed via Google Forms [7] through the authors' social networks, encompassing friends, colleagues, and family members. While this method facilitated rapid data collection, we acknowledge the potential for bias and sample imbalance due to the nature of the distribution channels.

Demographics: Participants were queried about their age, gender, education level, and occupation. This demographic information was crucial to understanding the contextual background of each respondent.

Lifestyle and Sleep Habits: The lifestyle section encompassed questions regarding exercise frequency, device usage, and screen time before sleep. In the sleep habits section, we focused on gathering detailed information about participants' sleep patterns. Key questions included:

- **Bedtime and Wake-up Time**: These questions helped calculate the actual duration of nighttime sleep, a vital metric for assessing sleep quality.
- **Sleep Quality**: Respondents rated their overall sleep quality, providing us with a subjective assessment of their sleep experience.
- **Sleep Onset**: We inquired about the time it typically takes for respondents to fall asleep, aiming to understand sleep onset difficulties.
- **Sleep Medication**: This question ascertained whether participants use any medication to aid their sleep, which is indicative of sleep quality issues.
- **Sleep Disturbances**: The frequency of disturbances such as waking up during the night or experiencing restless sleep was also captured.

Contributions: Our survey, though limited by its distribution method, provides valuable insights into sleep quality and its influencing factors. The subsequent statistical and machine learning analyses contribute to a nuanced understanding of sleep dynamics, laying the groundwork for future research in this field.

B. Preprocessing: Translation, Merging, and Cleanup

Translation: To harness the collected data effectively in the subsequent stages of analysis, it was imperative to amalgamate the survey responses from all language versions into a single dataframe. This posed significant challenges, as the survey questions and responses were presented in various languages, and even the English version contained verbose questions and answers. To address these issues, we:

1) Constructed a JSON file containing a translation mapping from the original column headers to abbreviated column headers in English.
2) Utilized the `rename` function from the `pandas` library to update the dataframe with the new headers [8].

For the translation of cell values, a similar approach was employed, where JSON files served as the translation mapping. These mappings were then applied to the dataframe using the `map` function to convert the cell values into concise English terms [9].

Merging & Cleanup: Once translations for both headers and cell values were standardized, we employed the `concat` function from `pandas` to stack the individual dataframes from each language version into a unified dataframe [10]. However, this merged dataset contained several inconsistencies, including:

- Heights recorded as less than 100cm, primarily in the Bengali version of the survey, where respondents preferred inches to centimeters. For these cases, heights were converted from inches to centimeters.
- Respondents' confusion between 24-hour and 12-hour time formats, resulting in miscalculations of sleep duration, with some reported durations nearing 20 hours per day. We addressed this by filtering out erroneous values and correcting the time format.

Furthermore, actual night sleep duration was calculated from the 'Bedtime' and 'Wake-up Time' responses. For respondents providing height and weight, Body Mass Index (BMI) was computed, offering additional metrics pivotal to the analysis. Following these corrections, the dataset was ready for subsequent stages of examination.

IV. Analysis

TABLE I
Descriptive statistics of the sleep study data

	Height (cm)	Weight (kg)	BMI	Sleep Quality	Night Sleep[2]
count	83	92	80	108	105
mean	165.31	67.42	24.55	3.44	7.04
std	8.32	12.80	4.25	0.82	1.37
min	150.00	43.00	18.52	2.00	1.67
25%	160.00	59.80	22.02	3.00	6.50
50%	167.00	68.00	24.69	3.00	7.00
75%	171.00	75.00	27.31	4.00	8.00
max	185.00	100.00	35.85	5.00	9.75

Descriptive statistics: As shown in Table I and II:

[2]Calculated night time sleep based on Bedtime and Wake-up time

TABLE II
DESCRIPTIVE STATISTICS OF SLEEP STUDY PARTICIPANTS

	Age Group	Gender	Education Level	Occupation	Exercise (days/week)
count	108	108	108	108	108
unique	5	3	4	7	4
top	25-34	Male	Master's	Student	1-2 Days
freq	72	67	47	47	43
	Device Usage (hrs/day)	Screen Time Before Sleep (hrs/min)	Sleep Onset (min)	Bedtime	Wake-up Time
count	108	108	108	104	108
unique	4	4	4	18	20
top	7+ Hours	30-60mins	15-30mins	23:00	07:00
freq	43	45	55	24	18
	Nap Duration (min)	Sleep Duration (hrs/24hr)	Sleep Disturbances	Sleep Medication	Language
count	108	107	108	108	108
unique	5	3	5	2	4
top	No Nap	6hrs+	Rarely	No	English
freq	61	64	48	48	68

- **Sleep quality**: Respondents rated their sleep quality around 3 on a 5-point scale, indicating moderate quality.
- **BMI**: The average Body Mass Index (BMI) is around 23.55, ranging from 16.5 to 39.4.
- **Night sleep duration**: The average night sleep duration is around 7 hours, with a wide range of 1.67 to 9.75 hours.
- **Age Group**: The most common age group among respondents is 25-34.
- **Gender**: A slightly higher number of male respondents compared to females.
- **Education Level**: The majority of respondents have a Master's degree.
- **Occupation**: Many respondents are students.
- **Exercise Days/Week**: '1-2 Days' is the most common response for exercise frequency.
- **Device Usage (hrs/day)**: A large portion of respondents use devices for '7+ Hours' per day.
- **Screen Time Before Sleep**: '30-60 Minutes' is the most common duration for screen time before sleep.
- **Sleep Disturbances**: 'Rarely' is the most frequent response, indicating that most respondents rarely experience sleep disturbances.
- **Sleep Medication**: The majority of respondents do not use sleep medication.
- **Language**: English is the most common language among respondents.

Distribution Analysis: The analysis of sleep quality and night sleep duration distributions, as depicted in Figures 1 and 2, reveals the following insights:

- **Sleep Quality Distribution:** The majority of the respondents report their sleep quality within the range of 2 to 4, with a concentration of responses at a sleep quality score of 3. Notably, a smaller proportion of participants indicate experiencing excellent sleep quality, as reflected by the fewer scores of 5.
- **Night Sleep Duration Distribution:** Observations from the histogram suggest a normal distribution of sleep duration, predominantly centered around the 7-hour mark. This finding is in line with standard sleep recommendations. Notably, the occurrences of extremely short (<5 hours) or long (>9 hours) sleep durations are relatively infrequent.
- **Gender-Based Sleep Differences:** The boxplots in Figure 2 indicate a striking similarity in sleep quality between male and female participants. Additionally, the comparison of night sleep duration across genders reveals closely aligned distributions, with minor variations primarily in the lower and upper quartiles.

These observations contribute to a nuanced understanding of the distribution patterns of sleep quality and duration among the survey participants, highlighting subtle differences and commonalities in sleep experiences.

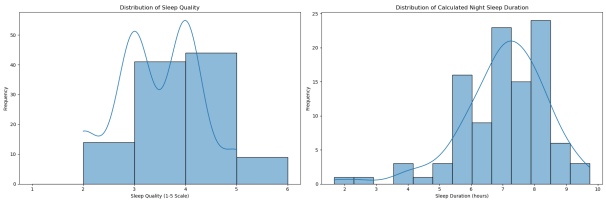

Fig. 1. Distribution of Sleep Quality and Duration

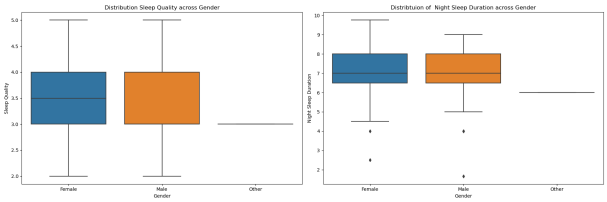

Fig. 2. Distribution of Sleep Quality and Duration across Gender

Correlation analysis: The heatmap in Figure 3 illustrates these following correlations:

- **Sleep Quality:** Strong negative correlation with Sleep Disturbances (-0.55), indicating better sleep quality is associated with fewer disturbances. Moderate negative correlation with Sleep Onset Time (-0.32), suggesting that quicker sleep onset is associated with better sleep quality.

- **BMI:** Slight negative correlation with Calculated Night Sleep Duration (-0.19), suggesting that higher BMI might be slightly associated with shorter sleep duration, although the relationship is weak. Moderate negative correlation with Device Usage (-0.28), indicating that higher BMI is associated with less device usage.
- **Calculated Night Sleep Duration:** Negative correlation with Age Group (-0.23), indicating that older age groups might have shorter sleep duration. All other correlations with Calculated Night Sleep Duration are weak.
- **Age Group:** Moderate positive correlation with BMI (0.31), suggesting that higher BMI values are more prevalent in older age groups.
- **Sleep Onset Time:** No significant correlations with other variables, aside from the moderate negative correlation with Sleep Quality.
- **Nap Duration:** Weak correlations with all other variables.
- **Exercise Days/Week:** Slight positive correlation with Calculated Night Sleep Duration (0.22), implying that more exercise might be related to slightly longer sleep duration. Weak correlations with all other variables.
- **Sleep Disturbances:** Aside from the strong negative correlation with Sleep Quality, Sleep Disturbances show weak correlations with other variables.
- **Device Usage (hrs/day):** Moderate negative correlation with BMI (-0.28), as previously mentioned. Weak correlations with all other variables. This heatmap indicates that while some variables are correlated, most relationships are weak. The strongest observed relationships involve sleep quality, particularly its negative correlation with sleep disturbances and sleep onset time. This suggests that variables affecting the quality of sleep have a more significant impact on sleep disturbances and the time it takes to fall asleep. The correlations involving BMI, age group, and device usage suggest demographic and behavioral patterns but are not strong enough to imply causation.

V. Hypothesis Testing

Overview

The observed correlations from the data analysis and visualizations suggest several hypotheses that could be explored through further analysis:

- *Impact of Sleep Disturbances on Quality:* Given the strong negative correlation between sleep disturbances and quality, we can hypothesize that increased sleep disturbances are likely to negatively impact the quality of sleep.
- *Age in Relation to Sleep Duration:* The negative correlation between age and calculated night sleep duration leads to the hypothesis that sleep duration may decrease with age.
- *Relationship Between Sleep Onset Time and Quality:* The moderate negative correlation observed between sleep onset time and quality suggests that a longer time to fall asleep might be associated with poorer sleep quality.
- *Influence of Exercise on Sleep Duration and Quality:* The slight positive correlation between exercise days per week and sleep duration hints at a potential hypothesis that increased physical activity could contribute to longer and possibly better quality sleep.
- *Nap Duration's Effect on Nighttime Sleep Duration:* Although the correlation is weak, we could investigate whether the duration of naps has any effect on the duration of nighttime sleep.

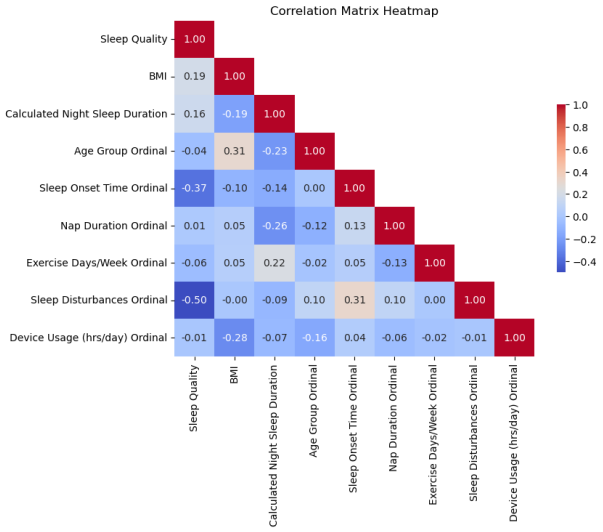

Fig. 3. Correlation Heatmap between variables

Hypothesis I - Increased Sleep Disturbances Negatively Impact Sleep Quality

Null Hypothesis (H_0): The level of Sleep Disturbances has no impact on Sleep Quality.

Alternative Hypothesis (H_1): The level of Sleep Disturbances negatively impacts Sleep Quality.

To investigate the relationship between sleep disturbances and sleep quality, we utilized Spearman's rank correlation due to its ability to measure monotonic relationships without assuming a linear relationship between variables, which is suitable for ordinal data. Additionally, the Chi-squared test was employed to evaluate the independence of categorical variables.

TABLE III
Statistical Tests for Sleep Disturbances and Sleep Quality

Test Type	Statistic	P-value	Conclusion
Spearman's ρ	-0.453	4.2×10^{-7}	Reject H_0, accept H_1
Chi-squared	46.03	6.85×10^{-6}	Reject H_0, accept H_1

The Spearman correlation coefficient of -0.453 suggests a moderate negative correlation, indicating that increases in sleep disturbances are associated with decreases in sleep

quality. Given the p-value of 4.2×10^{-7}, which is substantially below the alpha threshold of 0.05, we reject the null hypothesis, affirming the alternative hypothesis of a negative impact of sleep disturbances on sleep quality.

Similarly, the Chi-squared test yields a statistic of 46.03, surpassing the critical value for 12 degrees of freedom. Alongside a p-value of 6.85×10^{-6}, this provides strong evidence against the null hypothesis, suggesting a significant association between sleep disturbances and sleep quality in our dataset.

Hypothesis II - Older People Have Shorter Night Sleep

Null Hypothesis (H_0): Age group has no impact on Sleep Duration.

Alternative Hypothesis (H_1): Age group has a negative correlation with Sleep Duration.

The Pearson test was chosen for its ability to detect linear relationships between two continuous variables. Given the fairly normal distribution of Sleep Quality, it serves as an appropriate measure for this analysis. On the other hand, The Kendall's Tau test was selected for its robustness in assessing the strength of association between two ordinal variables, making it suitable for the nature of our age group data.

Pearson Correlations Test: The Pearson correlation coefficient was calculated as -0.232 with a p-value of 0.0088, indicating a weak negative correlation between age group and sleep duration. This suggests a trend where sleep duration may decrease slightly as age increases. Given the significance level of 0.05, the null hypothesis can be rejected in favor of the alternative.

Kendall's Tau Test: A Kendall's Tau test yielded a correlation of -0.204 with a p-value of 0.00619, reinforcing the presence of a weak negative correlation. The results allow us to reject the null hypothesis, affirming the alternative that there is a statistically significant, though mild, negative association between age group and sleep duration.

TABLE IV
STATISTICAL TESTS FOR AGE GROUP AND SLEEP DURATION

Test Type	Correlation	P-value	Conclusion
Pearson	-0.232	0.0088	Reject H_0, accept H_1
Kendall's Tau	-0.204	0.00619	Reject H_0, accept H_1

Both tests corroborate the alternative hypothesis, suggesting that as individuals progress into higher age categories, a decrease in sleep duration may occur. However, the correlation is not strong, pointing to the influence of other factors not accounted for in this analysis.

Hypothesis III - Longer Sleep Onset Time Worsens Sleep Quality

Null Hypothesis (H_0): Increase in Sleep Onset Time has no impact on Sleep Quality.

Alternative Hypothesis (H_1): Increase in Sleep Onset Time leads to a decline in Sleep Quality.

The relationship between Sleep Onset Time and Sleep Quality was analyzed using Spearman's rank correlation for its suitability with ordinal data and the Chi-squared test to assess variable independence.

TABLE V
STATISTICAL TESTING OF SLEEP ONSET TIME WORSENS SLEEP QUALITY

Test Type	Statistic	P-value	Conclusion
Spearman's ρ	-0.377641	2.80×10^{-5}	Reject H_0, accept H_1
Chi-squared	19.297 (DoF: 9)	0.022782	Reject H_0, accept H_1

The Spearman correlation coefficient of -0.377641 suggests a moderate inverse relationship, where longer Sleep Onset Time is associated with poorer Sleep Quality. The p-value of 2.80×10^{-5} indicates that this correlation is statistically significant and not due to chance.

Similarly, the Chi-squared test with a statistic of 19.297119 and a p-value of 0.022782 also suggests a significant association between Sleep Onset Time and Sleep Quality. Though not a strong relationship, it is sufficient to reject H_0 and accept H_1, indicating a notable connection that warrants further investigation.

Hypothesis IVa - Weekly Exercise Frequency Improves Sleep Quality

Null Hypothesis (H_0): Weekly exercise frequency has no impact on Sleep Quality.

Alternative Hypothesis (H_1): Weekly exercise frequency improves Sleep Quality.

Spearman's correlation was used to test for a potential positive association between exercise frequency and sleep quality.

TABLE VI
SPEARMAN CORRELATION AND CHI-SQUARED TEST FOR EXERCISE FREQUENCY AND SLEEP QUALITY

Test Type	Statistic	P-value	Conclusion
Spearman's ρ	-0.0577	0.7235	Accept H_0, reject H_1
Chi-squared	13.2199	0.1529	Accept H_0, reject H_1

The Spearman test resulted in a correlation coefficient of -0.0577 with a p-value of 0.7235, indicating no meaningful relationship between exercise frequency and sleep quality. The negative coefficient suggests a slight inverse trend, but the high p-value shows this is likely due to chance.

The Chi-squared test further supports the lack of evidence for an association. With a Chi2 Stat of 13.2199 and a p-value of 0.1529, the result does not exceed the critical value required to reject the null hypothesis at the 0.05 significance level.

Conclusion: The data does not support a significant relationship between weekly exercise frequency and sleep quality. Both Spearman's correlation and the Chi-squared test outcomes suggest that other unexplored factors might be more influential in determining sleep quality.

Hypothesis IVb - Weekly Exercise Frequency Improves Sleep Duration

Null Hypothesis (H_0): Weekly exercise frequency has no impact on Sleep Duration.

Alternative Hypothesis (H_1): Weekly exercise frequency improves Sleep Duration.

Spearman's correlation and ANOVA tests were conducted to evaluate the relationship between exercise frequency and sleep duration.

TABLE VII
STATISTICAL TESTS FOR EXERCISE FREQUENCY AND SLEEP DURATION

Test Type	Statistic	P-value
Spearman's ρ	0.1805	0.0327
ANOVA	2.9423	0.0367

The Spearman correlation coefficient of 0.1805 indicates a positive but weak relationship between exercise frequency and sleep duration, with a p-value of 0.0327 suggesting statistical significance.

The ANOVA test yields an F-statistic of 2.9423 and a p-value of 0.0367, confirming significant differences in sleep duration among different levels of exercise frequency.

Conclusion: While the Spearman test suggests a positive correlation between exercise frequency and sleep duration, the relationship is weak. The ANOVA test reveals significant differences in sleep duration across exercise frequency groups. These findings imply that while an increase in exercise frequency might be associated with longer sleep duration, further analysis is required to fully understand the dynamics of this relationship.

Hypothesis V - Increased Nap Time Decreases Night Sleep Duration

Null Hypothesis (H_0): Increase in Nap Time has no impact on Night Sleep Duration.

Alternative Hypothesis (H_1): Increase in Nap Time shortens Night Sleep Duration.

Spearman's correlation and Kendall's Tau tests were conducted to assess the relationship between nap duration and night sleep duration.

TABLE VIII
STATISTICAL TESTS FOR NAP DURATION AND NIGHT SLEEP DURATION

Test Type	Statistic	P-value	Conclusion
Spearman's ρ	-0.1208	0.1099	Accept H_0, reject H_1
Kendall's Tau	-0.096	0.1145	Accept H_0, reject H_1

The Spearman correlation coefficient of -0.1208 and a p-value of 0.1099 suggest a very weak negative relationship between nap duration and night sleep duration, but this is not statistically significant. Similarly, Kendall's Tau test shows a correlation coefficient of -0.096 and a p-value of 0.1145, also indicating a very weak negative correlation that does not provide enough statistical evidence to reject the null hypothesis. These results suggest that increased nap duration does not have a significant impact on reducing night sleep duration.

Conclusion of Hypothesis Testing

Our statistical analysis at a significance level of 0.05 has led to several noteworthy findings regarding the impact of various factors on sleep quality and duration:

- **Sleep Disturbance:** There is a significant negative correlation between sleep disturbances and sleep quality. This indicates that individuals with more sleep disruptions tend to have poorer sleep quality, underscoring the adverse effects of these disturbances.
- **Age Group and Sleep Duration:** A weak, yet statistically significant negative correlation exists between age group and sleep duration. This trend suggests that sleep duration may decrease slightly as age increases, highlighting the need for focused sleep health in older age groups.
- **Sleep Onset Time:** The analysis shows a significant relationship between sleep onset time and sleep quality. A quicker sleep onset is associated with better sleep quality, implying that faster sleep initiation is beneficial for overall sleep experience.
- **Exercise Weekly Frequency:** Our findings do not provide strong statistical evidence to suggest that increased exercise frequency significantly improves sleep quality. However, there is an observed correlation between higher exercise frequency and longer sleep duration, though its impact on sleep quality remains inconclusive.
- **Nap Duration:** The data does not establish a significant link between nap duration and night sleep duration. Therefore, it remains unclear whether napping habits substantially affect the total duration of nocturnal sleep, indicating an area for further research.

In conclusion, while certain factors like sleep disturbances and sleep onset time show clear correlations with sleep quality, other factors such as exercise frequency and nap duration present more complex relationships that require further exploration. The results underscore the multifaceted nature of sleep and its susceptibility to a range of influences, both clear and subtle.

VI. MODELING

Overview

In this modeling section, we focus on using variables that exhibited high correlations in our hypothesis testing, namely **Nap Duration**, **Exercise Days/Week**, **Sleep Disturbances**, and **Age Group**. These variables, identified as significant predictors of sleep patterns, are employed in various machine learning models to predict night sleep duration and sleep quality. The subsequent subsections detail the performance of each model, providing an analysis and comparison of their effectiveness in these prediction tasks.

A. Predicting Night Sleep Duration

The task of predicting night sleep duration was approached with multiple regression models. The models' performances

TABLE IX
REGRESSION MODELS: PERFORMANCE METRICS FOR PREDICTING NIGHT SLEEP DURATION

Model	Mean Squared Error	R^2 Score
Linear Regression	1.3389	0.2457
Decision Tree Regressor	3.4192	-0.9263
K-Nearest Neighbors	1.6241	0.0850
Support Vector Machine	1.5909	0.1037

were evaluated based on Mean Squared Error (MSE) and R-squared (R^2) scores.

The Linear Regression model showed the best performance with the lowest MSE, indicating higher predictive accuracy, and a modest R^2 score, suggesting a reasonable fit to the data. In contrast, the Decision Tree Regressor performed poorly, with a high MSE and a negative R^2 score, indicating overfitting. KNN and SVM had similar performance levels, but both were less effective than Linear Regression.

B. Predicting Sleep Quality

For sleep quality, classification models including Multiclass Logistic Regression, Decision Tree Classifier, Random Forest, KNN, and SVM were evaluated based on accuracy and weighted average F1-score.

TABLE X
CLASSIFICATION MODELS: PERFORMANCE METRICS FOR PREDICTING SLEEP QUALITY

Model	Accuracy	Weighted Avg F1-Score
Multiclass Logistic Regression	0.32	0.28
Decision Tree Classifier	0.55	0.62
Random Forest	0.55	0.62
K-Nearest Neighbors	0.50	0.47
SVM	0.59	0.52

In predicting sleep quality, the SVM model emerged as the most accurate with the highest overall accuracy and F1-score. Both Decision Tree and Random Forest models showed commendable performance but were slightly less effective than SVM. The Multiclass Logistic Regression model, however, had the lowest accuracy and F1-score, indicating its limited suitability for this particular task.

Modeling Summary

The modeling analysis indicates that for predicting night sleep duration, Linear Regression provides the most reliable results, while for sleep quality, SVM is the most effective. Although some models like the Decision Tree Regressor and Multiclass Logistic Regression showed limitations in predictive power, they still contribute valuable insights. Overall, each model's performance highlights the complexity of predicting sleep-related outcomes and underscores the potential for enhanced model development with more comprehensive data.

VII. CONCLUSION

Our study embarked on an in-depth examination of the complex interplay between various lifestyle factors and their impacts on sleep quality and duration. The significant correlations identified through hypothesis testing between sleep disturbances, age, exercise frequency, nap duration, and sleep onset time with sleep quality and duration highlight the intricate nature of sleep dynamics.

In the modeling phase, we utilized key variables, drawing on findings from hypothesis testing, to predict night sleep duration and sleep quality using various machine learning techniques. The regression models, especially Linear Regression, showed promising results in predicting night sleep duration, whereas the SVM classification model proved most accurate in predicting sleep quality. Nevertheless, these models demonstrated certain limitations, indicating areas for further refinement in future research.

This research contributes valuable empirical insights into the factors influencing sleep, suggesting that targeted lifestyle modifications could enhance sleep quality and duration. The study also underscores the potential of machine learning in sleep research, despite the challenges posed by the complexity of sleep patterns.

However, our study faced several limitations:

- The survey distribution mainly among friends and colleagues led to imbalanced data, potentially skewing the representation of broader populations.
- Time constraints limited our ability to fine-tune the models for optimal performance, which might have yielded more nuanced insights.
- Our limited expertise in the biological aspects of sleep may have constrained our ability to fully interpret the complex interactions between various factors and sleep patterns.

In conclusion, while providing valuable insights and contributing to the body of knowledge on sleep health, our study also paves the way for future research. It highlights the multifaceted influences on sleep and the importance of sophisticated analytical tools in understanding these complexities. The limitations noted herein can serve as guideposts for future studies, aiming to enhance our understanding and ultimately improve sleep health and overall well-being.

Future Work

The insights gained from our current study pave the way for several enhancements in future research endeavors. To further refine our understanding of sleep dynamics and improve predictive modeling, we propose the following areas of focus:

Refinement of Survey Methodology: Future iterations of the survey should involve collaboration with health professionals and reference established studies in health and sleep research. This approach will help in identifying new and important variables that could offer deeper insights into sleep patterns and their influencing factors. By grounding the survey in a more scientifically robust framework, we can ensure that it captures a comprehensive range of factors relevant to sleep quality and duration.

Addressing Data Imbalance: One of the key challenges in our study was the data imbalance resulting from the limited scope of survey distribution. To mitigate this in future research, the survey should be disseminated across a broader demographic. This includes reaching participants beyond the authors' immediate social circles to achieve a more representative sample. Efforts should be made to ensure class balance in the dataset, which would enhance the reliability and generalizability of the findings.

Advancements in Modeling Techniques: Given additional time and resources, fine-tuning the current models could significantly enhance their predictive capabilities. This process would involve more rigorous testing, validation, and optimization of parameters for models like Linear Regression and SVM, which showed promise in our study. Furthermore, exploring a diverse range of modeling techniques, including clustering algorithms and neural networks, could provide new perspectives and potentially more accurate predictions. These advanced methods could uncover complex patterns in the data that simpler models might miss, offering a richer understanding of the factors influencing sleep quality and duration.

In essence, the future work in this field should aim at not only addressing the limitations identified in our current research but also expanding the horizons of sleep studies through innovative methodologies and advanced analytical techniques. By doing so, we can contribute more effectively to the body of knowledge on sleep health and its critical role in overall well-being.

Acknowledgment

We would like to express our deepest gratitude to a number of individuals and organizations whose support was invaluable to the success of this project:

1) Our heartfelt thanks go to our families and friends. Their unwavering emotional support and active participation in our survey played a crucial role in this research. Their encouragement and involvement were key to the study's success.
2) We extend our sincere appreciation to Professor Eunhui Kim, Yuna Jeong, and Teaching Assistant Nilesh Kumar Srivastava at KISTI. Their profound expertise, guidance in analytical techniques, and consistent support throughout the semester have greatly enriched our learning experience and contributed immensely to our research.
3) A special thank you to our dedicated team members. This project was a collaborative effort involving brainstorming, designing the survey, data analysis, hypothesis formation, and modeling. The hard work, commitment, and collective efforts of each team member were instrumental in bringing this project to fruition.
4) We are grateful to the University of Science and Technology (UST) and the Korea Institute of Science and Technology Information (KISTI) for organizing this remarkable program. Their scholarship has provided us with a unique opportunity to learn and experience the technological advancements in Korea. Their support has been pivotal in our academic and professional growth.

This project stands as a testament to the power of collaboration, mentorship, and support, and we are thankful to everyone who contributed to its success.

REFERENCES

[1] Pearson correlation coefficient - wikipedia. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. (Accessed on 12/15/2023).
[2] Spearman's rank correlation coefficient - wikipedia. https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient. (Accessed on 12/15/2023).
[3] Chi-squared test - wikipedia. https://en.wikipedia.org/wiki/Chi-squared_test. (Accessed on 12/15/2023).
[4] T. Ohida, A. Kamal, M. Uchiyama, K. Kim, S. Takemura, T. Sone, and T. Ishii, "The Influence of Lifestyle and Health Status Factors on Sleep Loss Among the Japanese General Population," *Sleep*, vol. 24, no. 3, pp. 333–338, 05 2001. [Online]. Available: https://doi.org/10.1093/sleep/24.3.333
[5] B. Gómez-González, E. Domínguez-Salazar, G. Hurtado-Alvarado, E. Esqueda-Leon, R. Santana-Miranda, J. A. Rojas-Zamorano, and J. Velázquez-Moctezuma, "Role of sleep in the regulation of the immune system and the pituitary hormones," *Annals of the New York Academy of Sciences*, vol. 1261, no. 1, pp. 97–106, 2012. [Online]. Available: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2012.06616.x
[6] E. Preišegolavičiūtė, D. Leskauskas, and V. Adomaitiene, "Associations of quality of sleep with lifestyle factors and profile of studies among lithuanian students," *Medicina (Kaunas, Lithuania)*, vol. 46, pp. 482–9, 07 2010.
[7] Dreamharmony sleep survey. https://docs.google.com/forms/d/e/1FAIpQLSfJD8RAKNcRU6fJVxgMPSump2wiMIZPlH0QK3_wU6qwMxOEjw/viewform. (Accessed on 12/15/2023).
[8] (2023, Dec.) pandas.DataFrame.rename — pandas 2.1.4 documentation. [Online; accessed 16. Dec. 2023]. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rename.html
[9] (2023, Dec.) pandas.DataFrame.map — pandas 2.1.4 documentation. [Online; accessed 16. Dec. 2023]. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.map.html
[10] (2023, Dec.) pandas.concat — pandas 2.1.4 documentation. [Online; accessed 16. Dec. 2023]. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.concat.html