

Personal Wellness: A Data-Driven Approach to Stress and Sleep Prediction with Machine Learning Using Lifestyle and Health Variables

Nicholas-Paul A. Webster

Georgia Southern University

B.S. Computer Science

Abstract

Machine learning analysis of various lifestyle factors, the task of identifying and modeling relationships between behavioral variables and outcomes in personal wellness, has become more important with the introduction of technology in all facets of day-to-day life. The emergence of large-scale health datasets enables data-driven research to find insights within sleep and stress-patterns. Traditional approaches to studying the relationships between sleep and stress often have relied on simple statistical methodologies which might fail to consider how various variables may interact with the findings, such as caffeine intake, occupation, age, etc. This project aims to analyze related data and make predictions regarding a person's sleep duration and stress levels using supervised machine learning techniques applied to a Kaggle dataset by Jay Joshi containing over 15000 records. Linear Regression and Decision Tree models will be applied to model sleep-related outcomes as continuous variables, with stress levels serving as a critical behavioral factor. Results indicate that modern lifestyle features have limited effectiveness in predicting sleep duration but provide stronger predictive insight into sleep quality. This approach seeks to identify patterns between digital habits and personal wellbeing, while highlighting both the capabilities and limitations of machine learning methods in analyzing contemporary lifestyle data.

1. Introduction

Sleep and Stress are vitally important fundamental components of our individual health, each playing a critical role in both one's physical and mental well-being. Receiving the recommended amount of sleep supports cognitive function, emotional regulation, and physiological recovery, while having stress levels have been linked to an increased risk of disease and reduced quality of life. [1], [2].

With technology being integrated into all facets of our day-to-day life, it is possible that our modern-day lifestyle habits such as screen time, amount of time spent on the phone before bed, and caffeine intake, may significantly influence these relationships. Traditional statistical approaches have effectively captured insights into these effects but often fail to capture the

complex relationship between numerous modern lifestyle variables. With the implementation of Machine Learning techniques, there is a more versatile approach for analyzing such data and identifying patterns that may be overlooked using conventional methodologies.

This study analyzes a large-scale dataset comprising of data collected from over 15,000 individuals to investigate how modern lifestyle habits influence sleep-related outcomes. Linear Regression and Decision Tree models are applied to evaluate how features such as screen time and caffeine intake can be used to predict an individual's stress levels and sleep quality. Results indicate that the dataset features provide limited insight into predicting one's duration of sleep but are better suited to predicting the stress levels of an individual, highlighting the key potential and limitations of machine learning when it comes to modeling personal well-being.

2. Data and Methods

To conduct this study, a publicly available dataset from Kaggle containing over 15,000 records of individual lifestyle and behavioral metrics was utilized (Jay Joshi, 2023). The dataset includes various features such as age, gender, daily screen time, amount of time spent on the phone before sleep, sleep duration, sleep quality, stress levels, and caffeine intake. These variables were selected due to their representation of modern-day behavioral factors that can influence sleep and stress related outcomes.

Prior to modeling, the dataset was preprocessed to ensure sustainability and that it was suited for machine learning evaluation. Categories that were non-informative, such as the user ID, were removed to prevent bias and redundancy. The dataset was then divided into categorical and numerical features. The numerical features being age, daily screen time hours, minutes of phone usage before sleep, and caffeine intake. The categorical features, specifically gender and occupation, were transformed using one-hot-encoding to allow them to be incorporated into the model.

Although the dataset contained no missing values, imputation techniques were utilized in the preprocessing pipeline as a fallback mechanism to ensure consistency in our methodologies and compatibility with the modeling framework. This ensures that any missing numerical features were filled using the meaning of the respective column and that missing categorical features would be filled with the most frequent value. To guarantee that all numerical features equally contribute to model training, standardization was applied using z-score normalization via StandardScaler, which rescaled the features to have a mean of zero and a standard deviation of one.

Following preprocessing, the dataset was split into training and testing subsets with an 80/20 split with a fixed random state of 42 to ensure reproducibility. Then two supervised learning models were implemented: Linear Regression and Decision Tree. Linear Regression was used as a baseline model to evaluate the potential linear relationships between behavioral features and stress/sleep related outcomes. In contrast, the Decision Tree model was chosen due to its renowned ability to

capture non-linear relationships and interactions between variables without requiring explicit feature engineering.

Initial modeling approach was conducted by using sleep duration as the target variable, aligning with traditional measures of sleep health. However, the models that resulted from this demonstrated significantly weak predictive performance, with the output converging toward the mean and depicting a limited relationship between the selected features and total sleep duration. Based on this observation, the focus of the analysis was refined to be more geared toward predicting the quality of sleep, which provided a more meaningful and interpretable target variable, allowing the models to better capture relationships between behavioral factors and sleep outcomes.

Model performance was evaluated using multiple regression metrics, including Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics were chosen to provide a full assessment of prediction accuracy and model fit. The comparison of the models allows for an evaluation of both linear and nonlinear approaches in capturing the relationship between modern lifestyle behaviors and stress/sleep related outcomes.

To further improve model performance, cross-validation and basic hyperparameter tuning was performed using a grid search approach. A 5-fold cross-validation approach was used to evaluate model stability across multiple data splits, reducing reliance on a single train-test split. Additionally, key Decision Tree parameters, including maximum depth and minimum sample sizes, were adjusted to prevent overfitting and improve generalization. These steps helped ensure that the models captured meaningful patterns within the data rather than just noise.

3. Results

This study evaluated the relationship between various modern lifestyle factors and stress and sleep related outcomes using both statistical and machine learning approaches. Results are presented in two stages: an initial failed baseline model aimed at predicting sleep duration followed by a refined model predicting sleep quality. Following this, stress prediction results are presented along with conducted correlation analysis to support and interpret model performance.

3.1. Sleep Duration (Baseline Model)

The initial approach focused on creating a model for predicting sleep duration using lifestyle variables such as daily screen time, phone usage before sleep, stress level, and caffeine intake. However, the resulting model had an extremely poor predictive ability.

Linear regression generated a very low coefficient of determination ($R^2 = -0.004$), indicating that there was no relationship between the variables and the total hours of sleep. Prediction consistently converged at the mean, implying that behaviors have little effect on total hours of sleep. This finding aligns with earlier studies suggesting that total hours of sleep are dependent on other biological and environmental factors apart from observed behaviors. [2]

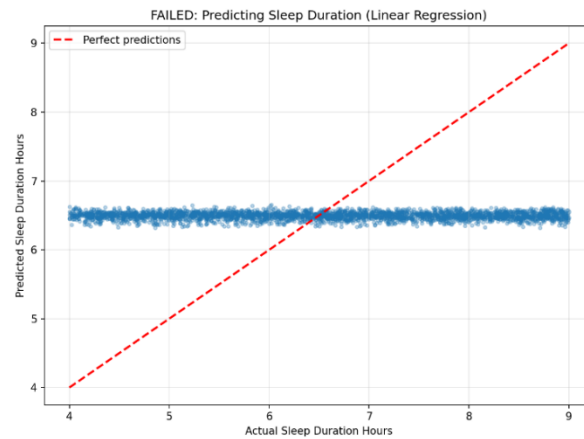


Figure 1. Predicted versus actual sleep duration values. Predictions clustered around the mean, showing poor predictive performance.

3.2. Sleep Quality and Stress (Refined Model)

Following the weak performance observed in training the baseline model, the study analysis was refined to be more geared towards sleep quality and stress levels as target variables. In contrast to the baseline results, both Linear Regression and Decision Tree demonstrated very strong predictive performance.

For predicting sleep quality, the Linear Regression model achieved an R^2 value of approximately 0.779, while the Decision Tree slightly outperformed with an R^2 score of 0.796. Similarly, predicting stress levels also depicted strong performance, with the Linear Regression model achieving an R^2 score of 0.8 and Decision Tree outperforming again with an R^2 score of 0.814. These scores indicate that the selected behavioral features are effective in modeling both sleep quality and stress level predictions.

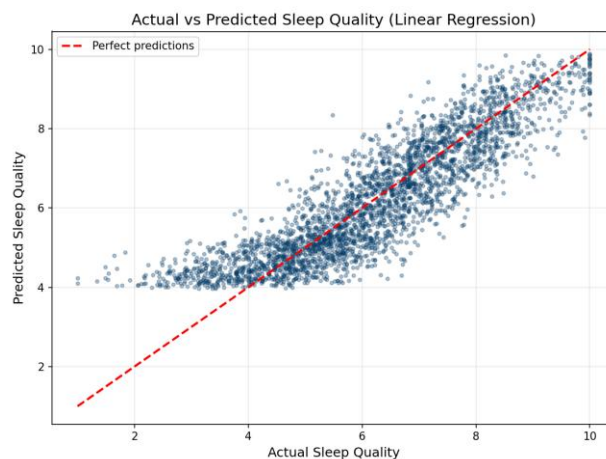


Figure 2. Predicted versus actual sleep quality score. Predictions align closely with actual values, indicating strong model performance.

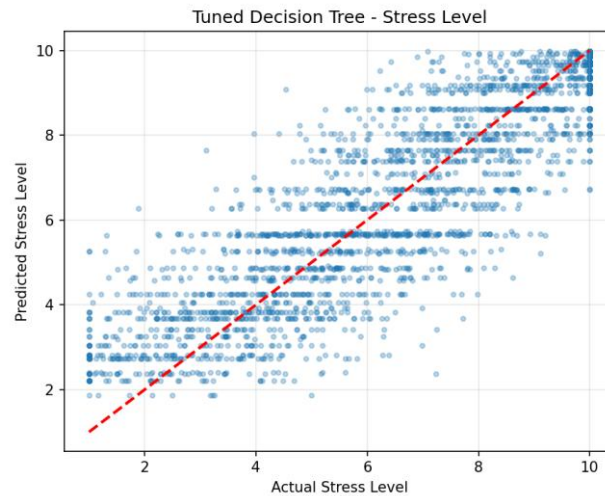


Figure 3. Predicted versus actual stress level values. Predictions show strong agreement with actual values, indicating that behavioral features effectively capture stress-related patterns.

3.3. Correlation Analysis.

Correlation analysis was conducted using the Pearson correlation coefficient to quantify relationships between numerical values. The results revealed several strong patterns.

Stress levels exhibited a strong negative correlation with sleep quality ($r = -0.857$), indicating that higher stress levels are associated with lower sleep quality. Similarly, the mental fatigue score ($r = -0.808$) and daily screen time ($r = -0.747$) also demonstrated a strong negative relationship with sleep quality.

In contrast, sleep duration showed a near-zero correlation with most behavioral variables, reinforcing the weak baseline performance discovered in the baseline model. This further supports the drawn conclusion that sleep duration is not directly influenced by modern lifestyle variables.

Finally, stress levels demonstrated a strong positive correlation with the mental fatigue score ($r = 0.949$) and daily screen time ($r = 0.879$), suggesting that increased fatigue and screen time engagement are associated with increased stress levels. These findings support the claim that modern lifestyle behaviors influence a person's stress, sleep, and overall personal well-being.

4. Discussion and Conclusion

The results of this study highlight both the strengths and limitations of using machine learning to model outcomes of personal wellness based on modern lifestyle factors. The study's initial failure to predict sleep duration effectively demonstrates that one's total sleep time is influenced by a combination of external factors that were not captured within the dataset. This reinforces the idea that sleep duration may not be easily modeled using observable lifestyle variables alone.

In contrast, the strong predictive performance observed when sleep quality and stress levels were modeled indicates that these targets were more heavily influenced by lifestyle factors. Features such as stress level, screen time, and time on the phone before sleep demonstrated meaningful relationships with sleep quality, suggesting that modern habits in this digital age play a large influence on sleep outcomes. Similarly, the ability to accurately predict stress levels further proves the connection that behavioral factors and one's personal well-being hold.

Correlation analysis supported these findings by revealing strong relationships between one's stress, mental fatigue, and sleep quality. The observed inverse relationship between stress and sleep quality shows that as stress increases, the quality of sleep will decrease. In addition to this, the strong association between screen time and stress reinforces the impact that prolonged screen time has on personal health [3].

Overall, this study shows that although machine learning models have their weaknesses when it comes to capturing some of the factors affecting human health, such as the overall sleeping time, they are very efficient when it comes to detecting patterns in more behavior-based aspects, such as the quality of sleep and stress. These findings showcase the potential that data-driven approaches alongside machine learning techniques hold in understanding personal wellness.

References

- [1] N. F. Watson et al.,
"Recommended amount of sleep for a healthy adult," *Sleep*, 2015.
- [2] F. P. Cappuccio et al.,
"Sleep duration and health outcomes," *Sleep*, 2010.
- [3] Cajochen et al., 2011
"Evening exposure to screens affects sleep and circadian rhythms."