Mihir Kalyanthaya
6/6/25
DataSci223 Project

I will be working with the Multilevel Monitoring of Activity and Sleep in Healthy people (MMASH) dataset. The researchers focus on stress-related disorders, sleep disturbances, and mood dysregulation which has placed a lot of stress on the healthcare system. They understand that traditional methods to assess these conditions rely heavily on retrospective self-reporting and intermittent clinical visits which are affected by multiple biases and do not capture real-time interactions between physiological and psychological states. The researchers aim towards developing proactive and continuous monitoring approaches that can detect early warning signs of mental and physical health deterioration. They also hope that ML can be used to multimodal wearable sensor data for predicting and monitoring physiological/psychological data. The researchers used wearable 24/7 activity trackers that collect data on continuous heart rate, triaxial accelerometer data, sleep quality, physical activity, and psychological characteristics (i.e., anxiety status, stress events and emotions), as well as saliva biomarkers (cortisol and melatonin). For this project, I want to explore the various associations between heart rate and sleep quality. The ultimate goal of their research was to identify modifiable predictors of poor sleep quality and enable early detection of vulnerable individuals for timely intervention. The MMASH dataset contains a wide range of psychophysiological variables over a 24-hour period. MMASH includes beat-to-beat heart rate interval data, wrist accelerometer readings, sleep fragmentation indices, self-reported stress and mood, behavioral activity logs, and salivary biomarkers. MMASH integrates continuous physiological monitoring with ecological momentary assessments of mood and stress. This multidimensional nature makes better for development of ML models for sleep quality prediction and stress monitoring which require temporal context and multimodal input data.

The dataset used for this study, Multilevel Monitoring of Activity and Sleep in Health People (MMASH), was collected through a collaborative effort between BioBeats and researchers at the University of Pisa. This dataset provides a multidimensional view of the psychophysiological states of 22 healthy young adult male participants over a 24-hour monitoring period. The dataset includes 22 male participants, aged between 20 and 40 years, with no reported physical or psychological illness. As such, the demographic distribution is limited—the population is homogeneous in terms of sex and health status, and no ethnic or socioeconomic data are provided. These limitations should be considered when generalizing findings to broader or clinical populations. The sample size is 22 patients with each contributing a full 24-hour record comprising of multiple temporal data streams. There are multiple key features in the dataset. The physiological data consists of inter-beat intervals, heart rate, and salivary hormone concentrations. Behavioral/Activity data has accelerometry data across x,y,z-axis, steps, and posture, as well as activity logs which include eating, screen usage, physical exertion, caffeine intake, and sleep. Sleep-related measures include objective sleep metrics (total sleep time, sleep efficiency, WASO,

and subjective sleep quality via the Pittsburgh Sleep Quality Index. The psychological state and stress features contain daily stress inventory (DSI), PANAS score (emotions at 5 timepoints), and STAI (state-trait anxiety inventory). The target variable is the PSQI which is the primary outcome for a regression task in which lower values indicate better perceived sleep quality. PSQI is labeled as such: values less than 6 mean good sleep quality and values greater than or equal to 6 indicate poor sleep quality. The researchers describe their questionnaires by the following. For the state anxiety survey, results range from 20 to 80. Scores less than 31 may indicate low or no anxiety, scores between 31 and 49 an average level of anxiety or borderline levels, and scores higher than 50 indicate a high level of anxiety or positive test results. The behavioral avoidance/inhibition index is a scale that measures the reinforcement sensitivity theory that establishes biological roots in personality characteristics, derived from neuropsychological differences. The BIS/BAS scales comprise a self-report measure of avoidance and approach tendencies that contains four sub-factors (a high score in one of the subscales describes the degree of that temperamental characteristic for the individual, according to the original sample). Bis facet reflects subject sensitivity toward aversive events that promote avoidance behaviors. Drive is individual persistence and motivational intensity. Reward is the propensity to show a higher degree of positive emotion for goal attainment. Lastly, 'Fun' is impulsivity and immediate reward due to sensory stimuli or risky situations. The DSI contains 58 self-reported measures which allow a person to indicate the events they experienced in the last 24 hours. After indicating which event occurred, they indicate the stressfulness of the event on a Likert scale from 1 (occurred but was not stressful) to 7 (cause me to panic). A higher total score value indicates increased frequency of perceived daily stress. Positive and Negative Affect Schedule (PANAS) gives a score rating between 5 and 50 for both positive and negative emotions. The higher the PANAS value, the higher is the perceived emotion. The researchers were interested in these daily activities: sleeping, laying down, sitting, studying, driving, moving, light/medium/heavy movement, screen usage, caffeinated drink consumption, smoking, drinking. The Actigraph contains accelerometer and inclinometer data recorded throughout the day. Lastly, they also have data on patient saliva which contains clock genes and hormones concentrations in the saliva before going to bed and after waking up. Two samples per participant were included, one before sleep and one after waking up. Melatonin levels are reported in μg of melatonin per μg of protein, while cortisol levels are in μg of cortisol per 100 μg of protein. The authors note that no clock genes and hormones concentrations data was provided for User_21 due to problems in the salivary samples. Some additional exploratory analyses that could be performed include temporal patterns of heart rate variability throughout the day and how activity levels influence it, correlation analysis between physiological measures and sleep quality, an association between cortisol/melatonin levels and sleep fragmentation, and developing a model testing three variables: sleep quality index, average heart rate, stress index (testing for confounding etc.).

	Variable	Description	Mean (SD)	Median [IQR]	Min–Max	n (missing)
0	Efficiency	Sleep efficiency (%)	83.91 (6.75)	85.22 [77.16–89.06]	73.49–94.23	22 (0)
1	Total Sleep Time	Total sleep time (min)	313.00 (84.31)	326.00 [253.50–342.75]	144.00–578.00	22 (0)
2	Number of Awakenings	Number of awakenings per night	19.27 (9.78)	18.50 [12.25–21.00]	4.00–44.00	22 (0)
3	Screen Time (min)	Screen time before bed (min)	52.91 (52.11)	36.00 [2.50–87.50]	0.00–163.00	22 (0)

Table 1. Descriptive statistics of key study variables (n=22)

The main problem of this research is a regression task with the goal of predicting perceived sleep quality metrics measured by PSQI based on physiological/behavioral/psychological inputs. I will also implement a multiclass classification task to categorize sleep quality. A binary classification task may also be defined using the standard threshold of PSQI ≥ 6 to distinguish between good and poor sleepers. Input features were derived from multimodal sources (physiology, activity, and emotion). Temporal feature extraction includes physiological signal processing heart rate variability features derived from RR intervals (RMSSD, pNN50, LF/HF ratio), circadian pattern features, integration of psychological assessments (temporal proximity of high stress scores to sleep periods and interaction terms between anxiety scores and physiological responses), and biomarker features (cortisol/melatonin ratios, rate of change between evening and morning samples). Dimensionality reduction will also be performed (PCA) to address multicollinearity in physiological metrics and feature selection using LASSO regression to identify predictive variables in linear associations. Random Forest regression and gradient boosted trees can be used to identify non-linear interactions. I could also perform time-series modeling after rigorous feature selection. For model selection, I will use 5-fold-cross validation with subject wise splitting for sample independence. For regression metrics, I will be using root mean squared error to quantify prediction error magnitude for sleep fragmentation index, mean absolute error which is less sensitive to outliers for robust assessment across individuals, r^2 to assess proportion of variance explained by the model, and concordance correlation coefficient to evaluate agreement between predicted and actual sleep quality values. Classification metrics will include balanced accuracy to deal with any class imbalance, F1 score which is the harmonic mean of precision and recall for each sleep quality class, AUC to evaluate discriminative ability between sleep quality categories, and a confusion matrix to identify misclassification patterns for clinical interpretations. Both regression and classification metrics allow for precise quantification of predicted sleep quality (regression) and clinical categorization (classification). RMSE provides a clinically

interpretable error magnitude in the units of the sleep fragmentation index, while balanced accuracy ensures fair evaluation across sleep quality categories. The Primary Clinical Endpoint is the prediction of sleep fragmentation index with RMSE < 10% of the population mean and classification of sleep quality categories with balanced accuracy > 0.80. This is best because sleep fragmentation directly impacts restorative value of sleep and has been linked to daytime fatigue and cognitive impairment, the dual regression/classification approach provides both precise measurements for research applications and interpretable categories for clinical decision support, and the threshold values (RMSE < 10%, balanced accuracy > 0.80) represent clinically meaningful performance that would substantively improve upon current subjective assessment methods. The endpoint will be evaluated on held-out test data with bootstrap confidence intervals to assess model robustness and generalizability.

This ML model would serve as a decision support tool integrated within personalized health applications (ex. wearable devices). The model would passively collect physiological and behavioral data over time and provide individualized predictions of perceived sleep quality, enabling early detection of sleep disturbances and stress-related health risks. Clinicians could use this information to monitor at-risk patients, initiate behavioral interventions, or prioritize follow-up visits.

The primary algorithm used in this analysis is multiple linear regression. This algorithm models the relationship between a continuous outcome variable (in this case, sleep efficiency) and multiple predictor variables (such as screen time, total sleep time, and number of awakenings). Multiple linear regression is appropriate for this problem because our main outcome is continuous, and our goal is to quantify and interpret the associations between several predictors and sleep efficiency. The key assumptions of linear regression are linearity between predictors and outcome, independence of errors, constant variance of errors (homoscedasticity), and normally distributed residuals. Linear regression performs well when these assumptions are met, and the predictors are not extremely collinear. The algorithm estimates coefficients by minimizing the sum of squared differences between observed and predicted values (the “least squares” method). Mathematically, the model is expressed as: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$ where Y is the outcome, X_1 to X_p are predictors, β_0 is the intercept, β_1 to β_p are coefficients, and ε is the error term. Ridge regression is a regularized version of linear regression that adds an L2 penalty to the loss function, shrinking the coefficients towards zero. This helps control overfitting, especially when predictors are highly correlated or when there are more predictors than observations. The penalty term is controlled by a tuning parameter (λ), which determines the extent of shrinkage. Ridge regression is appropriate when multicollinearity is present among predictors, as it can yield more stable estimates. The loss function minimized is: Sum of squared errors + $\lambda \times$ sum of squared coefficients. Lasso regression (Least Absolute Shrinkage and Selection Operator) is another regularized linear regression algorithm. It differs from ridge regression by using an L1 penalty, which can force some coefficients to be exactly zero, effectively performing variable selection. This is useful when it is suspected that only a subset of predictors are truly relevant. Like ridge, the amount of

regularization is controlled by a tuning parameter (λ). Lasso regression performs well in high-dimensional settings or when feature selection is desired but can be unstable if predictors are highly correlated. The loss function minimized is defined as: Sum of squared errors + λ × sum of absolute values of coefficients. Random forest regression is a non-parametric ensemble algorithm based on decision trees. It builds a large number of decision trees using bootstrapped samples of the data and averages their predictions to produce the final result. At each split in a tree, a random subset of predictors is considered, which helps to reduce correlation among trees and improve generalization. Random forest makes few assumptions about the data and can capture non-linear relationships and interactions between predictors, making it robust to outliers and capable of modeling complex patterns. However, the resulting model is less interpretable than linear models. Random forest works well when the relationship between predictors and outcome is non-linear or when there are complex interactions. Multiple linear regression was chosen as the primary algorithm due to its interpretability for the continuous outcome and question of interest. Ridge and lasso regression were used to assess the effects of regularization and to explore potential feature selection. Random forest regression was included to capture possible non-linear relationships and interactions among predictors, providing a robustness check for the linear models.

This project would be defined as an Exploratory Sleep Analysis. Our comprehensive analysis of sleep efficiency determinants (n=22) revealed several important findings. The number of awakenings emerged as the strongest predictor of sleep efficiency ($\beta = -0.421$, $p = 0.003$), with each additional awakening associated with a 0.42 percentage point decrease in sleep efficiency, consistent with established sleep medicine literature on the detrimental effects of sleep fragmentation. Contrary to prevailing hypotheses, screen time exhibited a modest positive association with sleep efficiency ($\beta = 0.020$, $p = 0.427$) across all models, although this relationship did not reach statistical significance. This counterintuitive finding persisted even after controlling for total sleep time and awakenings. Confounding analysis indicated that total sleep time substantially confounded the relationship between screen time and sleep efficiency, with the screen time coefficient decreasing by 17% when adjusting for sleep duration. The fully adjusted model demonstrated good explanatory power ($R^2 = 0.411$, Adj. $R^2 = 0.313$), with number of awakenings remaining the only statistically significant predictor. Bootstrap stability analysis (1,000 iterations) confirmed that while coefficient estimates showed considerable variability, the model's overall explanatory power was robust (82.9%). The model evaluation revealed limitations in predictive capability, with poor cross-validation performance ($R^2 = -2.69$) indicating overfitting—an expected outcome with small samples. Linear regression without interaction terms provided the most reliable predictions, suggesting the underlying relationships may be fundamentally linear despite the limited sample size. Diagnostics confirmed residuals appeared normally distributed (Shapiro-Wilk test: $W = 0.958$, $p = 0.448$), supporting the validity of our statistical approach. Since this problem is regression, I used r-squared, RMSE and RSE for metrics as the variable of interest (sleep efficiency) is a continuous variable. These findings should be interpreted cautiously given our small sample size.

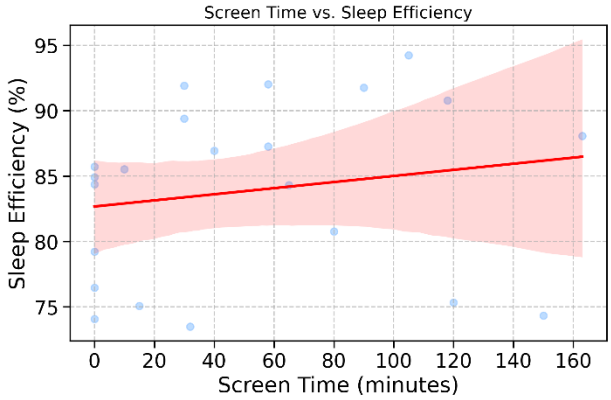
Additionally, I used a logistic regression model using “good” versus “poor” sleep efficiency as a binary outcome. The model included total sleep time, number of awakenings, and screen time as predictors. The results indicated that the number of awakenings had the strongest (borderline significant) association with lower odds of good sleep efficiency, while total sleep time and screen time were not significant predictors. I compared several approaches: linear regression for the continuous outcome, logistic regression for the binary outcome, and regularized models (ridge and lasso regression for both linear and logistic frameworks). Model fit and predictive performance were assessed using AIC, BIC, and cross-validated metrics such as R^2, accuracy, and AUC. The linear regression models exhibited poor fit, with negative cross-validated R^2 values, indicating worse performance than a null model. Logistic regression models showed modest performance, with a cross-validated accuracy of 0.58 and AUC of 0.60—barely above chance. Regularized models (ridge and lasso) did not improve upon these results. Overall, none of the models demonstrated strong predictive ability, likely reflecting the small sample size and limited signal in the predictors.

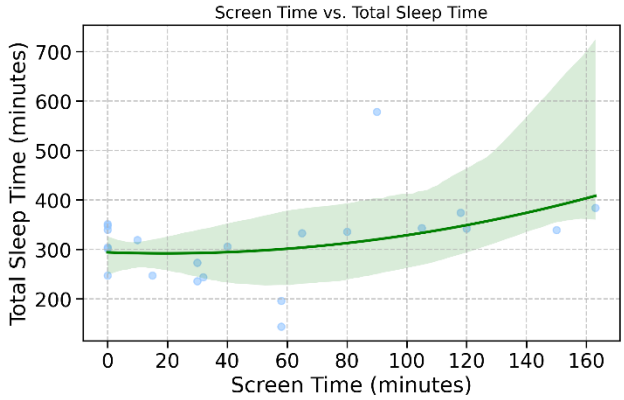
To further explore the associations between behavioral predictors and sleep efficiency, I applied two additional machine learning algorithms: a decision tree classifier and XGBoost. The decision tree classifier, using the number of sleep awakenings, total sleep time, and screen time as predictors, yielded a cross-validated accuracy of 0.47 (standard deviation = 0.18), indicating poor classification performance and substantial variability across folds. Visualization of the tree revealed that the number of awakenings was the primary splitting variable, with lower awakenings and reduced screen time modestly associated with higher sleep efficiency, though overall discrimination was limited. I then implemented an XGBoost classifier using the same predictors. XGBoost produced a cross-validated accuracy of 0.55 (standard deviation = 0.20) and an AUC of 0.48 (standard deviation = 0.08), suggesting performance only slightly better than random chance. Feature importance analysis from XGBoost highlighted sleep number of awakenings and total sleep time as the most influential variables, followed by screen time. However, the low predictive performance across both models indicates that, within this dataset, the selected behavioral factors do not provide strong or reliable discrimination between good and poor sleep efficiency. These findings emphasize the challenges of predictive modeling in small, limited datasets.

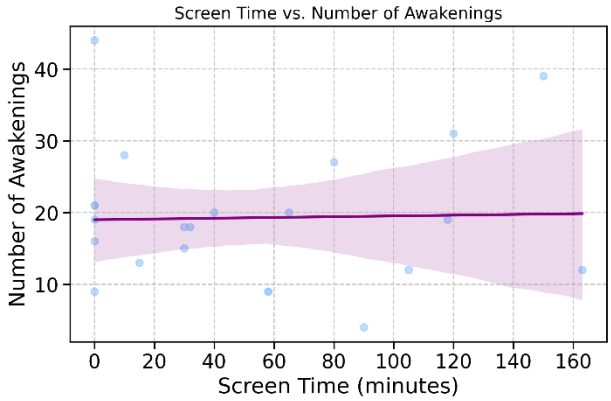
The unexpected positive association between screen time and sleep efficiency warrants further investigation with larger samples and more sophisticated measurement approaches. Future research should incorporate additional variables (e.g., chronotype, timing of screen exposure, light intensity) and longitudinal designs to establish causal relationships. This exploratory analysis generates valuable hypotheses for subsequent studies while highlighting the complex, multifactorial nature of sleep efficiency determinants.

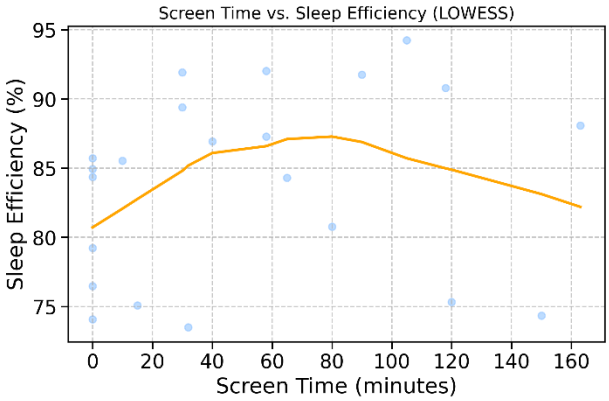
Though this model is overall comprehensive and fulfills the goal of this research, we need to be careful of biases present. The MMASH dataset is composed entirely of 22 healthy young adult males, introducing significant risks of selection bias and limited demographic representativeness. As such, the model’s predictions may not generalize well to other populations

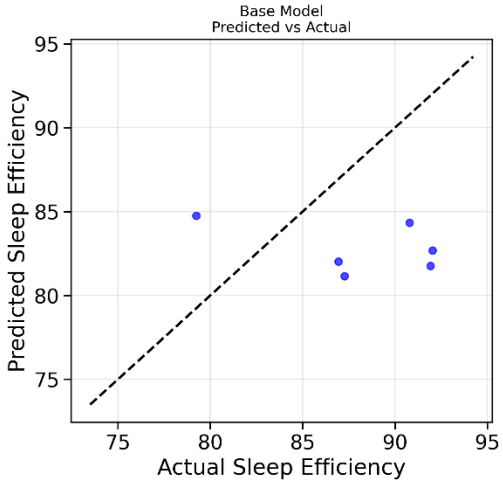
such as women, older adults, or individuals with chronic conditions. To mitigate this, we could expand the training dataset by collecting additional labeled data from diverse clinical cohorts (patients with insomnia, depression, PTSD) or implement reweighting or stratified sampling techniques to balance training data across demographic subgroups.

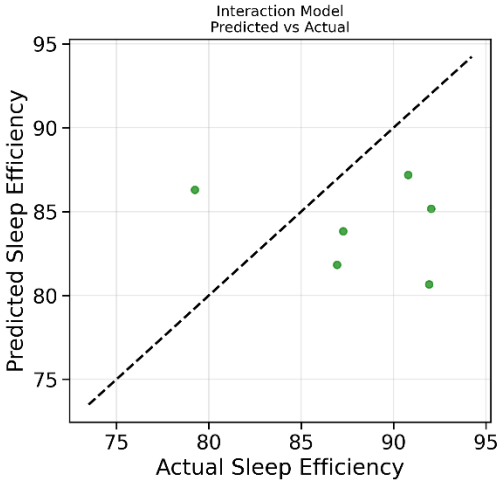


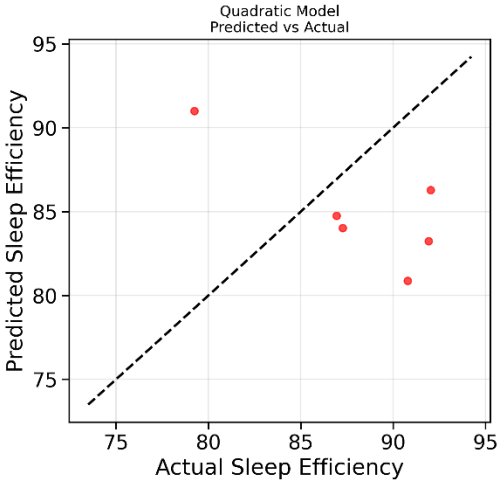












Citation:

- Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sirbu, A., Clifton, D., Martini, C., & Morelli, D. (2020). Multilevel Monitoring of Activity and Sleep in Healthy People (version 1.0.0). PhysioNet. https://doi.org/10.13026/cerq-fc86.
- Original Publication: Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sirbu, A., Clifton, D., Martini, C., & Morelli, D. (2020). A Public Dataset of 24-h Multi-Levels Psycho-Physiological Responses in Young Healthy Adults. Data, 5(4), 91. https://doi.org/10.3390/data5040091.