Sleep Habits and Lifestyle Factors

Project Midterm Report

Team Members: Nina, Lucas

Abstract

In this project, our goal is to determine whether machine learning can aid in predicting sleep disorders and the quality of sleep based on lifestyle and health factors. Our motivation for choosing this topic is that we found it an interesting area to learn about. We used two machine learning models: Logistic Regression and Linear Regression. So far, our preliminary results have matched our predictions, and the models are producing a high accuracy score.

1 Introduction

We obtained our dataset, “Sleep Health and Lifestyle Dataset,” from Kaggle. The data was collected by a Kaggle user a few years ago. We came across this dataset and became interested in the topic. After reviewing all the features and data listed, we believe machine learning could be applied to this data. Machine learning can help address this problem by analyzing the data we have so far to predict whether a person is likely to have a sleep disorder or not, based on their lifestyle and health factors. We will also try to predict their projected quality of sleep. This problem is significant because many people struggle with sleep, which can impact their day-to-day lives. At the end of this project, we will be able to suggest some ways to improve sleep quality and duration.

We applied two baseline algorithms aligned with our project objectives. Logistic Regression for predicting the presence of a sleep disorder (classification). Logistic regression is appropriate because the target variable is categorical (None, Insomnia, Sleep Apnea). It is interpretable, handles both continuous and categorical features, and provides coefficients that show which lifestyle or health factors most strongly influence the likelihood of a disorder. Then, we used linear regression to predict the quality of sleep score (1–10 scale). Linear regression is suitable for continuous outcomes and allows us to estimate the quantitative effect of predictors, such as activity level, stress, and blood pressure, on sleep quality.

Our expected outcome for this project is that increased levels of physical activity and daily steps, along with decreased levels of stress, blood pressure, and heart rate, will lead to a better quality of sleep and a reduced likelihood of sleep disorders. We do not believe that demographic factors, such as gender, age, and occupation, will have an impact on the outcome. So far, our preliminary results have aligned with our initial predictions.

2 Related Work

The first related work we found is a book from the Lecture Notes in Statistics series. This book, titled Linear Regression, is by Jürgen Groß. This reading provides insight into the theory and application of Linear Regression, which is one of the models we will be using.

The next related work we will reference is "An Introduction to Logistic Regression Analysis and Reporting," published by The Journal of Educational Research. The authors of this article are Joanne Peng, Kuk Lida Lee, and Gary M. Ingersoll. This article is interesting and relevant because it gives an overview of how logistic regression models are reported and analyzed. This will be very useful, as we

will be reporting on our logistic regression model for this project.

Lastly, we found an article titled "Analysis of Sleep Health and Lifestyle Factors: A Machine Learning Approach," written by Sourabh Sahu, Kalyani Tiwari, Garima Hardia, Rashmi Rani, Mohit Dhiman, and Amee Vishwakarma. This article closely relates to our project's objective and methods. This will be an interesting read and may provide us with tips on how to conduct our project.

All related works are cited in the References section.

3 SleepHealth-LR (Sleep Health Logistic–Linear Regression)

3.1 Overview

In this project, we employ two supervised machine learning models to explore the relationship between health and lifestyle factors and sleep outcomes: **Logistic Regression** to classify the presence of sleep disorders, and **Linear Regression** to predict an individual's self-reported sleep-quality score.

Both models are well-established baseline algorithms that allow for interpretable insights into which features most strongly influence sleep quality and disorders. Rather than proposing new architectures, our goal is to build a transparent, reproducible predictive pipeline using existing methods on a new dataset.

3.2 Dataset Collection

The dataset used in this study is the **Sleep Health and Lifestyle Dataset**, which is publicly available on Kaggle.

It contains approximately 400 survey entries collected from adults across different occupations and age groups.

Each observation includes:

- **Demographics:** Gender, Age, Occupation
- **Lifestyle factors:** Sleep Duration, Physical Activity (min/day), Daily Steps, Stress Level (1–10)
- **Physiological metrics:** Heart Rate (bpm), BMI Category, Blood Pressure (systolic/diastolic)

The target variables are:

1) **Sleep Disorder** (None, Insomnia, Sleep Apnea) for classification, and
2) **Quality of Sleep** (1–10 scale) for regression.

3.3 Data Pre-Processing

Data preprocessing was critical to ensure the models received clean and meaningful inputs.
We performed the following steps:

1. **Missing Values** — All rows containing missing target labels were dropped.
   For the feature matrix, missing numeric values were imputed using the median, while missing categorical values were imputed with the most frequent value.

2. **Feature Encoding** — Categorical variables, such as Gender, Occupation, and BMI Category, were one-hot encoded using scikit-learn, ensuring that unseen categories would not break the model.

3. **Feature Scaling** — Numeric attributes (e.g., age, steps, stress level, heart rate) were standardized to center and normalize distributions before regression.

4. **Train/Test Split** — The data was split 80/20 into training and test sets.
   For classification, the split was stratified by sleep-disorder category to preserve class proportions.

5. **Blood Pressure Parsing** — The raw "120/80" text field was separated into systolic and diastolic numeric features, improving interpretability and numeric modeling performance.

6. **Pipeline Integration** — All preprocessing steps were encapsulated in a **ColumnTransformer** inside an scikit-learn **Pipeline**, ensuring identical transformations during training and inference.

3.4 Feature Selection

Because this dataset is moderate in size and each variable has an intuitive meaning, we opted to include all features rather than performing automatic feature elimination.

This approach enables us to later inspect the model coefficients and identify which health and lifestyle factors are most predictive of the outcome.

We did, however, remove the **Person ID** column to prevent information leakage from a non-predictive identifier.

3.5 Model Design and Hyperparameters

- **Logistic Regression:**
  Used for categorical prediction of sleep disorders.
  We applied scikit-learn's **LogisticRegression** with the default L2 regularization.

  This solver efficiently handles both continuous and one-hot-encoded categorical inputs.

  Logistic Regression was chosen because it is interpretable (coefficients represent feature importance in log-odds) and provides a clear baseline for future, more complex classifiers such as Random Forest or SVM.

- **Linear Regression:**
  Used for continuous prediction of sleep-quality score (1–10).

  The model assumes a linear relationship between predictors and the response variable.

While simple, it provides a good initial benchmark and yields coefficients that directly quantify the contribution of each lifestyle factor to predicted sleep quality.

Future versions could extend this with Ridge or Lasso regression to handle multicollinearity.

No manual hyperparameter tuning was conducted at this stage, as the focus of the midterm report is on establishing a reliable preprocessing and modeling foundation rather than optimization.

3.6 Evaluation Metrics

Each task used metrics appropriate to its output type:

- **Classification (Logistic Regression).**
  We computed **Accuracy, Precision, Recall, F1-score**, and the **Confusion Matrix**.
  Accuracy summarizes overall correctness; the confusion matrix reveals the performance for each class.

- **Regression (Linear Regression).**
  We evaluated **R^2** (Coefficient of Determination) to measure explanatory power, **RMSE** (Root Mean Squared Error) to assess prediction spread, and **MAE** (Mean Absolute Error) for intuitive error magnitude on the 1–10 scale.
  A residuals-versus-predicted plot confirmed the absence of strong bias or heteroscedasticity.

3.7 Implementation Summary and Results

All experiments were implemented in Python 3.12 using pandas and scikit-learn.

The preprocessing and modeling pipelines ensured full reproducibility.

After training:

- **Logistic Regression Results**
  - Accuracy = **0.87**
  - Precision = Recall = F1 = 0.87 (per class)
  - Confusion Matrix = [[13, 2], [2, 14]], indicating a balanced detection of both Insomnia and Sleep Apnea.

    The top-weighted features included **Blood Pressure** (140/95, 135/90), **Heart Rate**, **Sleep Duration**, and **Occupation type** (e.g., Nurse or Salesperson). Elevated blood pressure and heart rate were positively correlated with sleep disorders, while longer sleep duration and active occupations were negatively correlated.

- **Linear Regression Results**
  - R^2 = **0.935**
  - RMSE = **0.31**

- MAE = **0.083**

These values indicate an excellent fit, explaining approximately 93% of the variance in the sleep-quality score with minimal prediction error.

The residual plot showed errors centered near zero, confirming the stability of the regression model.

3.8 Insights and Reflection

Both baseline models achieved strong predictive performance and meaningful interpretability.
The classification pipeline demonstrates that health and occupational variables can effectively distinguish between major sleep-disorder categories, while the regression pipeline reveals that stress level, blood pressure, and physical activity have measurable quantitative effects on sleep quality.
Although these simple linear models perform well, they may still under-represent nonlinear relationships (e.g., interactions between stress and activity).
Future work will explore regularized or tree-based methods (such as Ridge and Random Forest) to validate robustness and generalization on unseen data.

4 Preliminary Results

The models produced strong and reliable outcomes across both prediction tasks.

For the sleep-disorder classification, the Logistic Regression model achieved an overall accuracy of **0.87**, with precision, recall, and F1-score all equal to **0.87** on the held-out test set. The confusion matrix [[13, 2], [2, 14]] shows that both classes—Insomnia and Sleep Apnea—were predicted with balanced performance, misclassifying only four of the thirty-one test observations. This consistency indicates that the features supplied roughly equal discriminative information for both disorders. Examination of the model coefficients revealed that **Blood Pressure**, **Heart Rate**, and **Sleep Duration** were among the most influential predictors. Individuals with higher blood pressure or an elevated resting heart rate were more likely to be classified as having a sleep disorder, whereas a longer average sleep duration was associated with healthier sleep patterns.

For the sleep-quality regression, the Linear Regression model achieved an **R^2 of 0.935**, an **RMSE of 0.31**, and an **MAE of 0.083**, demonstrating an excellent linear fit that explains about 93 percent of the variance in the self-reported quality-of-sleep scores. On a 1-to-10 scale, the average prediction error is therefore less than one-third of a point, suggesting high predictive accuracy. The residuals plot showed that errors were distributed symmetrically around zero, confirming that the model exhibited no major bias or heteroscedasticity.

Overall, these preliminary results validate the modeling framework and preprocessing pipeline. Even without hyperparameter tuning or feature selection, both models successfully captured the major relationships between health and lifestyle indicators and sleep outcomes. The performance of these interpretable baselines establishes a strong foundation for future experimentation with regularized linear models, such as Ridge and Lasso, or nonlinear ensemble methods, like Random Forest and Gradient Boosting, which may capture higher-order interactions and further enhance predictive generalization.

5 Future plan

Building on the successful baseline results achieved so far, the next stage of the project will focus on enhancing model robustness, interpretability, and generalization.

Although both baseline models performed exceptionally well—achieving 0.87 accuracy for classification and an R^2 of 0.935 for regression—these results were obtained using relatively simple linear methods.

To verify that such high performance is not driven by linear assumptions or feature redundancy, we plan to experiment with regularized regression techniques (Ridge and Lasso) and nonlinear ensemble models such as Random Forest and Gradient Boosting.

These models will enable us to capture potential nonlinear interactions between lifestyle, physiological, and demographic factors that may influence sleep outcomes.

Additionally, we will perform cross-validation to ensure that model performance is stable across multiple random splits and not overfitted to the current train-test partition.

We also intend to implement feature-importance analyses (e.g., permutation importance and SHAP values) to better quantify the contribution of each predictor and validate whether the most influential features—such as blood pressure, heart rate, and sleep duration—remain consistent across models.

Finally, the regression and classification pipelines will be refactored for better reproducibility and deployment, enabling the generation of automated predictions and visual outputs (such as importance plots and correlation heatmaps) for inclusion in the final report.

By the end of the project, our objective is to deliver a validated, interpretable, and generalizable model that can reliably predict both the presence of sleep disorders and sleep quality from everyday health and lifestyle indicators—contributing to broader insights into behavioral and physiological factors that affect human sleep.

6 References

Groß, J. (2003). *Lecture Notes in Statistics: Linear Regression*. Springer.

Peng, J., Lida Lee, K., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*. https://doi.org/10.1080/00220670209598786

Sahu, S., Tiwari, K., Hardia, G., Rani, R., Dhiman, M., & Vishwakarma, A. (2025). Analysis of Sleep Health and Lifestyle Factors: A Machine Learning Approach. *Communications on Applied Nonlinear Analysis*, (Vol. 32 No. 2 (2025)). https://doi.org/10.52783/cana.v32.6132

Kaggle**.** *Sleep Health and Lifestyle Dataset.* Kaggle, n.d. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.