Sleep Habits and Lifestyle Factors

Final Report

Team Members: Nina De Grandis, Lucas Sorge

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. Our preliminary results matched our predictions, and the models are producing a high accuracy score. This is consistent with our final results.

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. Our preliminary results aligned with our initial predictions.

Through this project, our goal is not only to build accurate models but also to gain a deeper understanding of how lifestyle choices impact sleep and to highlight potential ways people can improve their sleep habits.

2 Related Work

We found and referenced many related works during this project. 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. Additionally, 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.

Next, we found many papers that give more insight to the methods we used. 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. “Random Forests” by Leo Breiman, 2001, and “Understanding Random Forests: From Theory to Practice” by Gilles Louppe in 2014 detailed the random forest method which we used often in this analysis because linear analysis had constraints that didn’t allow for anything we wanted to do. For the linear analysis, the papers, “Ridge Regression: Biased Estimation for Nonorthogonal Problems” by Arthur E. Hoerl & Robert W. Kennard (1970), and “Regression Shrinkage and Selection Via the Lasso” by Robert Tibshirani (1996), were both very helpful to provide a deeper insight into these methods. Lastly, we referred “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection” by Ron Kohavi in 1995 to compare models and see what we wanted to use for this project.

All of these related works proved useful when reflecting upon this project. They provided valuable insights on the details and comparisons of certain models. This was very helpful when deciding what models were best to use for this project.

All related works are cited in the References section.

3 Methods

This section describes the modeling approach used to investigate how lifestyle, physiological, and demographic variables contribute to sleep disorders and overall sleep quality. We detail the dataset preprocessing pipeline, feature engineering decisions, model classes, hyperparameter strategies, and evaluation metrics used across both prediction tasks: (1) classifying Insomnia vs. Sleep Apnea, and (2) predicting sleep quality on a 1–10 scale.

3.1 Dataset Collection and Structure

The dataset consists of 374 adults with measurements covering four domains:

- **Demographics:** Gender, Age, Occupation
- **Lifestyle factors:** Sleep Duration, Physical Activity, Daily Steps
- **Physiological measures:** Resting Heart Rate, Stress Level, Blood Pressure
- **Outcomes:** Sleep Disorder (categorical), Quality of Sleep (numeric)

A critical characteristic of the dataset is that **only 155 individuals have Sleep Disorder labels**, almost perfectly balanced between Insomnia (77) and Sleep Apnea (78). The remaining 219 unlabeled rows are still usable for regression but must be dropped for classification.

3.2 Data Preprocessing

Because the dataset contains a mix of numeric and categorical variables, a unified preprocessing pipeline was constructed using scikit-learn's ColumnTransformer. The pipeline ensures consistent handling during training, testing, and cross-validation.

3.2.1 Handling Missing Values

- Numeric variables → Median imputation
- Categorical variables → Most-frequent imputation

Median imputation was chosen for stability against skewed variables such as Daily Steps and Heart Rate.

3.2.2 Feature Engineering

Several important preprocessing steps were applied:

Blood Pressure Parsing

The original "Blood Pressure" variable was stored as a string (e.g., "130/85").

It was split into two meaningful numeric features:

- BP_Systolic
- BP_Diastolic

This transformation enabled the models to treat blood pressure as separate physiological contributors to sleep outcomes.

Dropping Non-Informative Identifiers

Person ID was removed because:

1. It contains no behavioral or physiological information.
2. It was nearly perfectly correlated with Age due to how the dataset was constructed.

One-Hot Encoding

Categorical features one-hot encoded:

- Gender
- Occupation
- BMI Category

To avoid multicollinearity, we used drop="first" so that each categorical variable is represented by (k – 1) encoded features.

Feature Scaling

All numeric variables were standardized using z-scores to ensure that linear models (Logistic Regression, Ridge, Lasso) treat all predictors fairly regardless of original units.

3.3 Train-Test Splitting Strategy

The dataset was divided using an 80/20 split wiht:

- **Stratification** for classification, ensuring balanced representation of both disorders in training and testing.
- **Random splitting** for regression, since the target is continuous.

This resulted in:

- **Classification:** 124 training cases, 31 test cases
- **Regression:** 299 training cases, 75 test cases

These splits support fair evaluation and avoid overfitting to the small labeled subset.

3.4 Feature Selection Rationale

Although formal feature selection techniques (e.g., RFE, mutual information) were not applied, exploratory analysis and correlation structure guided decisions:

- **Sleep Duration, Stress Level, Heart Rate, and BP values** were strongly correlated with sleep quality and were retained in full.
- **Daily Steps and Physical Activity** were kept due to domain relevance despite being correlated ($r \approx 0.77$).
- **Occupation categories** were left in because early analysis revealed unexpected predictive strength (e.g., nurses showing higher apnea risk).
- **Person ID** was removed due to non-informative correlation with Age.

The philosophy was to retain potentially meaningful predictors and let models (especially nonlinear ones) evaluate their importance.

3.5 Machine Leaning Models

To examine both linear and nonlinear relationships in the data, we implemented a diverse suite of models grouped into three categories.

3.5.1 Baseline Linear Models

Logistic Regression (Classification Task)

Used to separate Insomnia vs. Sleep Apnea.

- L2 regularization (ridge penalty)
- Solver optimized for small to medium datasets (liblinear)
- Coefficients used for interpretability analysis

This model provides insight into which features increase or decrease the likelihood of each disorder.

Linear Regression (Quality-of-Sleep Prediction)

Serves as the simplest baseline for regression.

We selected this model because:

- EDA showed nearly linear relationships between key predictors and sleep quality.
- Residual analysis confirmed no major linearity violations.

3.5.2 Regularized l.inear Models

Regularization helps prevent overfitting, especially when categorical variables expand dimensionality after one-hot encoding.

Ridge Regression (L2)

Penalizes large coefficients but keeps all predictors.

Well-suited for correlated features such as activity measures and cardiovascular variables.

Lasso Regression (L1)

Can force coefficients to zero, effectively performing variable selection.

Useful for understanding whether sleep quality depends on a sparse subset of features.

Hyperparameter α was tuned using grid search.

3.5.3 Nonlinear Ensemble Models

Linear models cannot capture interactions such as:

- "High stress + high heart rate" combinations
- Threshold effects (e.g., sleep duration < 6 hours)
- Non-monotonic patterns in activity and occupation

Therefore, we include Random Forests.

Random Forest Classifier

Advantages:

- Captures nonlinear boundaries between Insomnia vs. Sleep Apnea
- Handles interactions naturally
- Less sensitive to feature scaling
- Robust to noise and correlated predictors

Random Forest Regressor

Chosen because:

- Sleep quality may depend on branching, nonlinear relationships

- Feature importances provide interpretability
- Random forests excel on medium-sized mixed-feature datasets

Hyperparameter tuning included number of trees, depth, and minimum split size.

3.6 Hyperparameter Tuning

All tuned models used **5-fold cross-validation** on the training set.

- Logistic Regression: C ∈ {0.01, 0.1, 1, 10}
- Ridge Regression: α ∈ {0.1, 1, 10, 100}
- Lasso Regression: α ∈ {0.001, 0.01, 0.1, 1}
- Random Forests: n_estimators ∈ {100, 200}, max_depth ∈ {None, 10, 20}

Cross-validation was crucial because the classification dataset is small (n = 155), and performance on a single split may be misleading.

3.7 Evaluation Metrics

Classification Metrics

- **Accuracy**
- **Precision, Recall, F1-score**
- **Confusion Matrix**

These metrics highlight balance between detecting Insomnia vs Sleep Apnea.

Regression Metrics

- **R^2 (variance explained)**
- **RMSE (root mean squared error)**
- **MAE (mean absolute error)**

Together, they quantify both relative and absolute prediction error.

Model Interpretability Tools

To understand predictions and extract meaningful sleep-health insights, we used:

- Logistic regression coefficients

- Random Forest feature importances
- Correlation heatmaps
- Boxplots and distribution plots
- Residual analysis

These tools form the basis for the Discussion and Conclusion sections.

3.8 Summary of Modeling Approach

Our methodological framework emphasizes:

1. **Consistency** — Shared preprocessing ensures fair model comparison.
2. **Diversity** — Both linear and nonlinear models reveal different aspects of the data.
3. **Interpretability** — Coefficients and feature importances translate results into actionable behavioral insights.
4. **Robustness** — Cross-validation and regularization safeguard against overfitting.
5. **Domain Alignment** — Choices reflect known physiological drivers of sleep health (stress, BP, HR, duration).

This integrated approach positions us to evaluate not only prediction performance but also the underlying lifestyle and physiological patterns that shape sleep quality and disorder risk.

4 Experimental Results

This section summarizes the performance of all models evaluated for predicting (1) Sleep Disorder (classification) and (2) Quality of Sleep (regression). We compare baseline linear models with regularized variants and tree-based models, using an 80/20 train–test split and 5-fold cross-validation for tuning.

4.1 Baseline Models

Logistic Regression (Classification Baseline)

The baseline logistic regression model achieved:

- Accuracy: 0.839
- Macro-F1: 0.839

The confusion matrix shows relatively balanced performance across Insomnia and Sleep Apnea, each predicted correctly in ~85% of cases. Coefficient analysis revealed clinically meaningful patterns: shorter sleep duration and sedentary occupations were linked to insomnia, while higher heart rate, diastolic BP, and overweight status were associated with sleep apnea.

Linear Regression (Regression Baseline)

Baseline linear regression achieved:

- R^2: 0.967
- RMSE: 0.225
- MAE: 0.119

These metrics reflect the strong linear correlations in the dataset. Residual plots showed no systematic bias, suggesting the linear model is an appropriate starting point.

4.2 Regularized Linear Models

Ridge Regression

- Best α: 10
- Test R^2: 0.964
- RMSE: 0.233

Ridge closely matched the baseline model, indicating that the dataset does not suffer from severe multicollinearity and that most predictors meaningfully contribute to explaining sleep quality.

Lasso Regression

- Best α: 0.01
- Test R^2: 0.955
- RMSE: 0.261

Lasso performed slightly worse, suggesting that forcing sparsity removes signal. Sleep quality appears influenced by a broad combination of lifestyle and physiological factors rather than a small subset of dominant features.

Regularized Logistic Regression (C = 0.01)

- CV Accuracy: 0.887
- Test Accuracy: 0.774

Although cross-validation improved, test accuracy decreased due to small sample size. Insomnia recall remained high (93%), while Sleep Apnea recall dropped (62%), illustrating limitations of linear boundaries for this task.

4.3 Tree-Based Models

Random Forest Classifier

The Random Forest classifier achieved the strongest classification results:

- Accuracy: 0.903
- Macro-F1: 0.903

Performance was balanced across both disorders (14/15 Insomnia and 14/16 Apnea correctly classified). Feature importance highlighted Sleep Duration, Stress Level, Heart Rate, and BP as key predictors. The improvement over logistic regression indicates that disorder classification depends on nonlinear feature interactions.

Random Forest Regressor

The Random Forest regressor yielded the best regression results:

- R^2: 0.982
- RMSE: 0.164
- MAE: 0.046

Its superior performance suggests that sleep quality depends on nonlinear effects (e.g., stress thresholds, heart-rate variability) that linear models cannot fully capture.

4.4 Summary of Comparisons

Classification Performance

Model	**Accuracy**	**F1**
Logistic Regression	0.839	0.839
Reg. Logistic (C=0.01)	0.774	0.774
Random Forest Classifier	0.903	0.903

Regression Performance

Model	R^2	**RMSE**	**MAE**
Linear Regression	0.967	0.225	0.119
Ridge Regression	0.964	0.233	0.167
Lasso Regression	0.955	0.261	0.181
Random Forest Regressor	0.982	0.164	0.046

4.5 Insights

- **Nonlinear models outperform linear ones** in both tasks, showing that sleep outcomes arise from complex interactions among stress, heart rate, activity, and BMI.

- **Linear models remain highly interpretable**, highlighting clinical relationships such as stress → reduced sleep quality and elevated heart rate → increased likelihood of sleep apnea.
- **Random Forest models provide the best predictions overall**, demonstrating strong generalization despite the modest dataset size.

5 Conclusion

Overall, this project shows that machine learning is an effective tool for predicting sleep disorders and estimating sleep quality based on lifestyle and health factors. Our linear models helped confirm clear relationships in the data, such as the negative impact of high stress, elevated heart rate, and poor blood pressure on sleep outcomes. While these models were useful and easy to interpret, the nonlinear Random Forest models performed the best in both classification and regression tasks. Their strong results suggest that sleep patterns are influenced by more complex interactions that linear models cannot fully capture.

In the future, expanding the dataset, adding more detailed health or wearable data, and testing additional machine learning models could help improve prediction accuracy even further. Despite the limitations of the current dataset, our findings highlight the potential of machine learning to provide valuable insights into sleep health and to support better lifestyle recommendations moving forward.

6 References

Breiman, L. (2001). *Random forests*. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1145).

Louppe, G. (2014). *Understanding random forests: From theory to practice* (arXiv:1407.7502). arXiv.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

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

Peng, J., Lee, K. L., & 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, 32*(2). https://doi.org/10.52783/cana.v32.6132

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