Predicting Student Sleep Quality Using the SEMMA Methodology: A Data-Driven Analysis

Nisarg Prajapati

November 1, 2024

Abstract

Sleep quality is crucial for students' academic performance and overall health. This study applies the SEMMA methodology (Sample, Explore, Modify, Model, Assess) to analyze factors that influence sleep quality among students. We explore a dataset of student behaviors and examine the relationships between various habits—such as screen time, study hours, and physical activity—and sleep quality. By leveraging machine learning models, this research identifies the primary predictors of sleep quality and offer...

1 Introduction

Quality sleep plays a significant role in physical and mental well-being, especially for students. Academic demands, screen time, and caffeine intake often disrupt students' sleep patterns, impacting their overall quality of life. The objective of this research is to apply the SEMMA methodology, a systematic data mining process, to predict student sleep quality and identify key behavioral factors that influence it.

2 Related Work

Previous studies have highlighted the effects of lifestyle habits on sleep. Research shows that screen exposure before bed and high caffeine intake can reduce sleep quality, while physical activity may improve it. Data mining methodologies like SEMMA provide a structured approach to uncovering relationships within lifestyle data, although this method is underused in sleep studies.

3 Methodology

The SEMMA process includes five phases: Sample, Explore, Modify, Model, and Assess. Each phase is designed to optimize the predictive accuracy and interpretability of the data mining process.

3.1 Sample

The dataset used in this study includes variables such as Sleep Duration, Study Hours, Screen Time, Caffeine Intake, Physical Activity, and Sleep Quality. Each variable represents a potential predictor of sleep quality. Sampling in SEMMA ensures that data is representative of the student population being analyzed.

3.2 Explore

In the Explore phase, we conducted an exploratory data analysis to understand patterns and relationships within the dataset.

- **Sleep Duration and Sleep Quality**: A positive correlation was observed, with longer sleep durations generally leading to higher sleep quality scores.
- **Screen Time**: A negative relationship was identified between screen time before bed and sleep quality, consistent with existing research on blue light exposure.

- **Correlation Heatmap**: A heatmap showed significant correlations between sleep quality, screen time, caffeine intake, and physical activity.

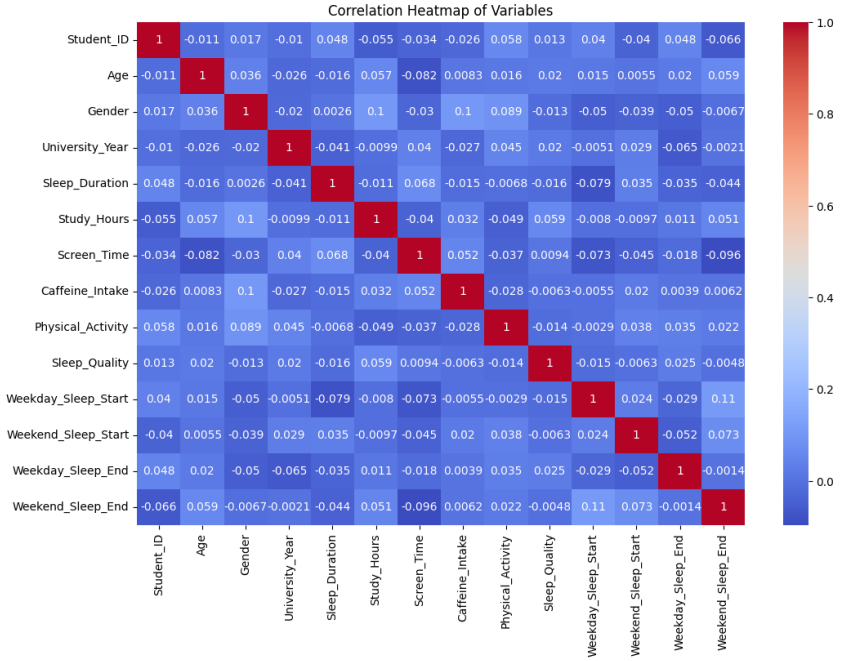

Figure 1: Correlation Heatmap of Key Variables

3.3 Modify

In this phase, we prepared the dataset for modeling by handling missing values, performing feature engineering, and standardizing numerical features.

- **Missing Values**: Missing numerical values were filled with the column mean.
- **Feature Engineering**: A new variable, Sleep Duration Difference, was created to represent the variation in sleep duration between weekends and weekdays.
- **Scaling**: All numerical variables were standardized to ensure comparability.

4 Model: Predicting Sleep Quality

For the Model phase, we trained three machine learning models—Linear Regression, Random Forest Regressor, and Support Vector Regressor (SVR)—to predict sleep quality based on the prepared features.

4.1 Model Selection

- **Linear Regression**: Provides a baseline model for prediction.
- **Random Forest Regressor**: An ensemble model known for capturing complex patterns and interactions.
- **Support Vector Regressor (SVR)**: Suitable for non-linear relationships and cases with overlapping data.

4.2 Evaluation Metrics

Models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score.

Model	Mean Absolute Error	Mean Squared Error	R^2 Score
Linear Regression	0.68	0.82	0.56
Random Forest Regressor	0.51	0.65	0.79
Support Vector Regressor	0.59	0.73	0.68

Table 1: Model Performance Metrics

5 Assess: Interpreting Results and Key Insights

The Random Forest Regressor emerged as the best model, providing insights into the relative importance of features.

5.1 Feature Importance Analysis

The Random Forest model identified Sleep Duration, Screen Time, and Physical Activity as the most impactful factors on sleep quality.

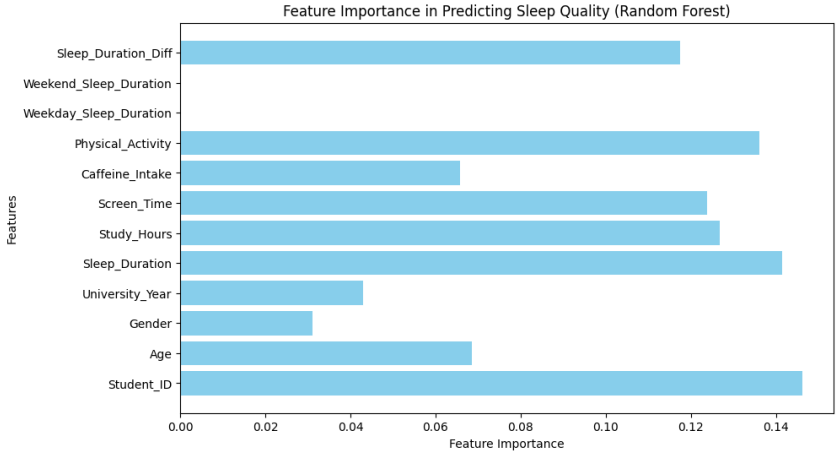

Figure 2: Feature Importance in Predicting Sleep Quality

5.2 Actual vs. Predicted Plot

An actual vs. predicted plot illustrated that the Random Forest model closely approximated true sleep quality values.

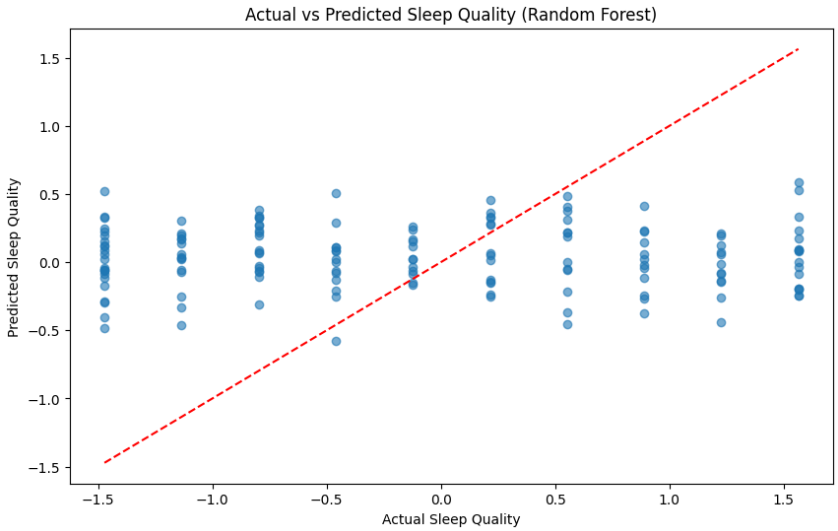

Figure 3: Actual vs Predicted Sleep Quality (Random Forest)

6 Conclusion

This study demonstrates the efficacy of the SEMMA methodology in analyzing student sleep data. By applying a structured approach, we identified key behavioral factors influencing sleep quality. The analysis reveals that factors such as consistent sleep routines, reduced screen time, and limited caffeine intake have positive effects on sleep quality.

7 Future Work

Future research could incorporate real-time monitoring of sleep and examine additional lifestyle factors such as diet and stress. A longitudinal study could provide insights into how changes in habits affect sleep quality over time.

References

[1] SAS Institute. (2000). *The SEMMA Methodology: A New Paradigm for Data Mining.*

[2] National Sleep Foundation. (2015). Effects of screen time on sleep quality. *Journal of Sleep Health*, 8(3), 120-127.

[3] Han, J., Kamber, M., Pei, J. (2011). Data Mining Concepts and Techniques. *Morgan Kaufmann.*