ADVANCED DATA ANALYSIS OF SLEEP TIME DATA

ST 3082 Data Analysis Project 2

Prepared by:
Group 14

- S16025 – Viruna Fernando
- S16368 – Poornima Tharangani
- S16305 – Pasindu Gamage

Content

List of Figures

List of Tables

Abstract

This report presents an advanced analysis of sleep time patterns using machine learning algorithms, building on insights obtained from a preliminary descriptive analysis. The study leverages a comprehensive sleep time dataset sourced from Kaggle to develop a predictive model aimed at forecasting sleep quality and duration. The model provides valuable insights for individuals, healthcare professionals, and researchers interested in sleep health and well-being. By employing predictive modeling techniques, our study enhances the understanding of factors affecting sleep quality and offers practical tools for improving sleep habits. This approach facilitates more personalized sleep interventions, ultimately contributing to better health outcomes and more informed decision-making in sleep management.

Introduction

Sleep plays a fundamental role in maintaining overall health and well-being, yet modern lifestyles often interfere with the quality and quantity of sleep individuals get. In this context, optimizing sleep patterns becomes essential for better health outcomes and enhanced daily performance. This report explores the intricate dynamics of sleep time prediction, with the goal of constructing a predictive model that forecasts sleep duration based on key daily activities. Using a dataset that includes predictors such as relaxation time, work hours, workout time, caffeine intake, reading time and phone usage, this study aims to identify the factors that most significantly impact sleep. The model developed here will serve as a valuable tool for individuals and health professionals to customize daily activities, promoting optimal sleep and improving overall well-being. By predicting sleep time, the report offers actionable insights that can help individuals adjust their routines for healthier sleep patterns, thereby fostering better decision-making in lifestyle management.

Description of the Question

Sleep is a critical factor in maintaining overall health, cognitive function, and physical performance. However, modern lifestyles, including long work hours, excessive screen time, caffeine consumption, and irregular exercise habits, often disrupt sleep patterns. Understanding the relationship between daily activities and sleep duration is essential for individuals who rely on optimal rest to perform at their best.

Many individuals, such as bodybuilders, athletes, long-haul drivers, healthcare professionals, students, and corporate employees, depend on sufficient and high-quality sleep for peak performance and well-being. Despite growing awareness of sleep hygiene, personal lifestyle factors make it challenging to predict and optimize sleep time effectively. This study aims to address this gap by constructing a predictive model to forecast sleep duration based on key daily activities, including caffeine intake, phone usage, reading time, relaxation time, work hours, and workout time.

By identifying the factors that most significantly influence sleep, this model serves as a valuable tool for individuals looking to improve their sleep hygiene. Healthcare professionals, fitness trainers, researchers, and those in high-performance professions can benefit from these insights to develop personalized sleep strategies. The ability to predict sleep time and adjust daily routines accordingly fosters better health management, enhanced cognitive function, and improved physical recovery.

Therefore, our primary objectives are:

- To construct a reliable model to predict sleep time based on daily activities.
- To provide insights into managing daily routines for optimal sleep.
- To help individuals and professionals make informed lifestyle adjustments to enhance sleep quality and overall well-being.

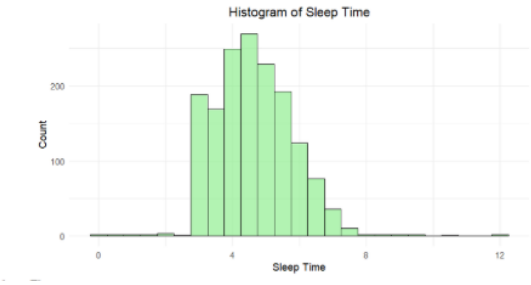
Description of the Dataset

The **"Sleep Time Prediction"** dataset is designed to analyze the relationship between various daily activities and sleep duration. The dataset originally contained **2000 observations and 7 variables**; however, **44 observations were removed** due to unrealistic and unusual data, resulting in a final dataset of **1956 observations and 6 variables**. The dataset includes key factors such as reading time, phone usage, workout duration, work hours, and caffeine intake, all of which influence sleep time. This dataset serves as a valuable resource for researchers, health professionals, athletes, bodybuilders, and individuals looking to optimize their daily activities for better sleep.

No.	Variable	Type of Variable	Comments
1	**Sleep Time**	Numerical-Continuous	Total sleep duration measured in hours. This is the response variable.
2	**Reading Time**	Numerical-Continuous	Time spent reading per day, measured in hours. Reading can promote relaxation and improve sleep quality.
3	**Phone Time**	Numerical-Continuous	Time spent using a phone per day, measured in hours. Excessive screen exposure before bedtime may reduce sleep duration.
4	**Workout Time**	Numerical-Continuous	Total duration of physical exercise per day, measured in hours. Regular exercise can influence sleep quality and duration.
5	**Work Hours**	Numerical-Continuous	Total work duration per day, measured in hours. Long work hours may decrease available sleep time.
6	**Caffeine Intake**	Numerical-Continuous	Total caffeine consumption per day, measured in milligrams (typically in three-digit values). High caffeine intake can negatively affect sleep duration, especially if consumed later in the day.

Table 1: Description of variables

Important Results of the Descriptive Analysis



The analysis revealed that the **majority of individuals sleep between 4 to 6 hours**, falling under the **short sleep** category. Based on standard sleep health classifications, sleep durations of less than 7 hours are considered **insufficient**, whereas **7-9 hours** is optimal for overall well-being.

Figure 1: Distribution of sleep Time

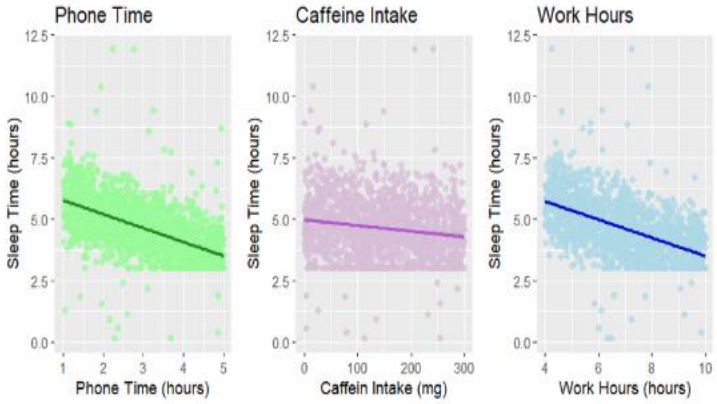


Figure 2: Scatterplot of phone Time, Caffeine Intake and Work Hours

Phone Usage and Sleep Duration

The scatter plot illustrates a **clear negative relationship** between **phone usage and sleep time**. As phone usage increases, **sleep duration declines noticeably**. The presence of **outliers** suggests that while most individuals experience reduced sleep with higher phone usage, some are less affected. This supports existing research on the **detrimental effects of excessive screen time before bed**, which can interfere with melatonin production and delay sleep onset.

Caffeine Intake and Sleep Duration

While the relationship between **caffeine intake and sleep time** is **not as strong** as phone usage or work hours, a **negative trend is still apparent**. The scatter plot indicates that individuals consuming high amounts of caffeine tend to have **shorter sleep durations**. However, some individuals with **moderate caffeine intake** appear to **sleep better** compared to extreme users or complete non-users. This suggests that **individual tolerance levels** and **timing of caffeine consumption** may play a role in sleep disruption.

Work Hours and Sleep Duration

A **strong negative relationship** is evident between **work hours and sleep duration**. The scatter plot highlights that individuals working **8+ hours per day** tend to experience the **sharpest decline in sleep time**. This aligns

with findings that longer work hours contribute to **chronic sleep deprivation**, potentially leading to fatigue, decreased productivity, and long-term health issues.

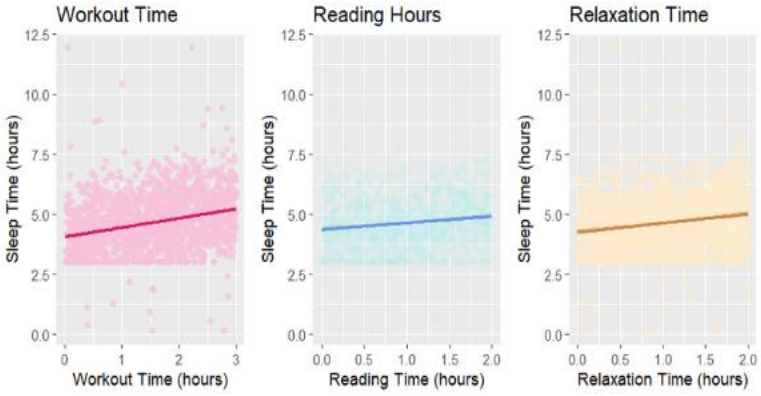


Figure 3:Scatterplot of Workout Time, Reading Hours and Relaxation Time

Workout Time and Sleep Duration

The scatter plot suggests a **slightly positive association** between **workout time and sleep duration**. Individuals who engage in **more workout time** tend to have **longer sleep durations**. However, the presence of **outliers** indicates that some individuals who exercise more actually experience **shorter sleep**, possibly due to **intense late-night workouts** disrupting sleep patterns. This highlights the importance of **exercise timing** in maintaining a healthy sleep schedule.

Relaxation Time and Sleep Duration

Higher **relaxation time** appears to have a **weak but positive correlation** with sleep duration. While the **trend suggests** that individuals who allocate more time for relaxation tend to **sleep longer**, the relationship is **not as strong** as phone usage, work hours, or caffeine intake. This could be due to **variations in relaxation activities**, where some methods (e.g., meditation) may enhance sleep quality more than others (e.g., passive screen time).

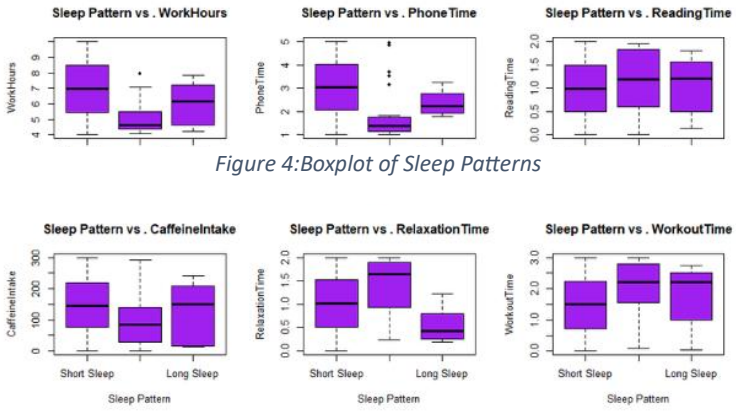


Figure 4:Boxplot of Sleep Patterns

Reading Time and Sleep Duration

Unlike other factors, **reading time does not show a clear association** with sleep duration. However, a **subtle trend** suggests that individuals who **spend more time reading** tend to **sleep slightly longer**. This might indicate that reading, especially before bedtime, serves as a **calming activity**, promoting better sleep compared to screen-based entertainment.

The relationship between sleep patterns and various lifestyle factors reveals significant trends. Individuals with short sleep durations (less than 7 hours) tend to have the highest work hours, with a high median and wide variability, whereas those with optimal sleep (7-9 hours) exhibit moderate work hours, suggesting a balanced lifestyle, and long sleepers (more than 9 hours) work significantly fewer hours with lower variability.

Phone usage follows a similar pattern, with short sleepers displaying the highest usage and widest spread, optimal sleepers maintaining moderate and controlled screen time, and long sleepers having the lowest phone usage. Reading time increases with sleep duration, with short sleepers reading the least, optimal sleepers engaging in more balanced reading habits, and long sleepers dedicating the most time to reading. Caffeine intake is highest among short sleepers, likely as compensation for insufficient rest, moderate for optimal sleepers, and lowest for long sleepers, suggesting less reliance on stimulants.

Relaxation time is lowest for short sleepers, likely due to demanding work hours and excessive phone usage, higher for optimal sleepers, and highest for long sleepers, who prioritize rest and recovery. Finally, workout time is least among short sleepers, possibly due to fatigue, peaks in optimal sleepers, indicating a well-balanced routine, and is moderate in long sleepers, falling between the other two groups.

These patterns suggest that optimal sleep duration supports a more balanced lifestyle with controlled screen time, moderate caffeine consumption, adequate relaxation, and the highest engagement in physical activity.

Important Results of the Advanced Analysis

Before conducting the advanced analysis, we performed **data preprocessing** to ensure data quality and suitability for modeling. Our dataset originally contained **2,000 observations and 7 predictors**, with **no missing values or duplicates**.

To identify **unusual observations**, we first examined whether any time-related variables exceeded the **24-hour range**, but no such values were found. However, upon calculating the **total time allocation per row**, we identified **44 observations** exceeding **24 hours**. Since these were **unrealistic**, we removed them from the dataset before proceeding with further analysis.

Outlier Detection and Principal Component Analysis (PCA)

- We **identified some outliers** in the **sleep time response variable**. However, removing them **did not significantly affect the fitted model**, so they were retained.
- A **Principal Component Score Plot** was generated to understand the distribution of observations in the latent space defined by **Component 1 (Comp 1) and Component 2 (Comp 2)**.
- Since **data points were tightly packed around the origin**, this indicated **no strong clustering or grouping** in the first two components.
- **Key PCA Findings:**
    - **Workout Time** and **Caffeine Intake** had **large positive loadings** on **Component 2**.
    - **Work Hours** had a **large negative loading** on **Component 1**.

- The response variable (**Sleep Time**) had a **strong positive loading** on **Component 1**.
- The first two components explained only **33% of the variance** in the predictors, suggesting that **additional components** are needed to capture the full structure of the data.

Multiple Linear Regression

As the first option, we select the Multiple Linear Regression since MLR serves as a strong baseline model before moving to complex models. The results of descriptive analysis ensure that the predictor variables are not correlated to each other. So, fitting a multiple linear regression may be reasonable. We fitted a multiple linear regression model with all the predictors "Workout Time", "Reading Time", "Phone Time", "Work Hours", "Caffeine Intake", "Relaxation Time".

We calculated Train RMSE value, Test RMSE value, Train R square and Test R square values and the results are given below,

	RMSE	R square
Training set	0.5800	0.7569
Testing set	0.5523	0.7782

Table 2:Evaluation metrics for MLR

The test RMSE and train RMSE are closer, so the model may generalize well. Test R square value significantly high which means that the model explains approximately 80% of the total variation. Overall model accuracy is good according to the results we obtained.

Since the MLR is an assumption-based model, we conducted model adequacy check for the linearity assumption, Multicollinearity, Independence and Homoscedasticity of residuals and multivariate normality.

Linearity Assumption

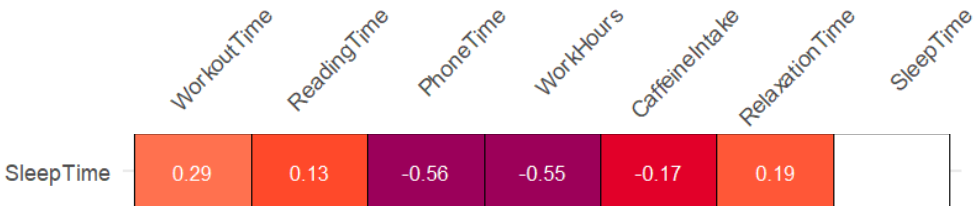


Figure 5: Correlation values for predictor variables

Correlation plot underscores small associations between sleep time and predictors except the Phone Time and Work Hours. So, it indicates that assuming a linearity of the model may be reasonable, but the strength of the linearity is limited.

Multicollinearity

The VIF table ensures that all the VIF values are less than 10, indicating no multicollinearity among predictors.

Independence and Homoscedasticity

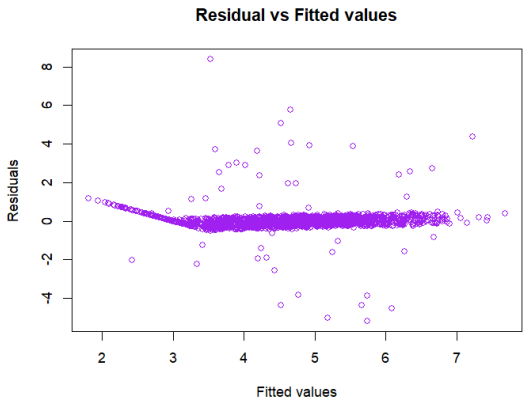


Figure 6:Residual Vs Fitted Values for MLR

Most of the points are scattered around zero mean level and no any systematic pattern in the points. It indicates that the Independence and Homoscedasticity assumptions are held.

Normality assumption

- **Shapiro Wilk Test**: Test statistic = 0.4745

  p-value < 2.2e-16

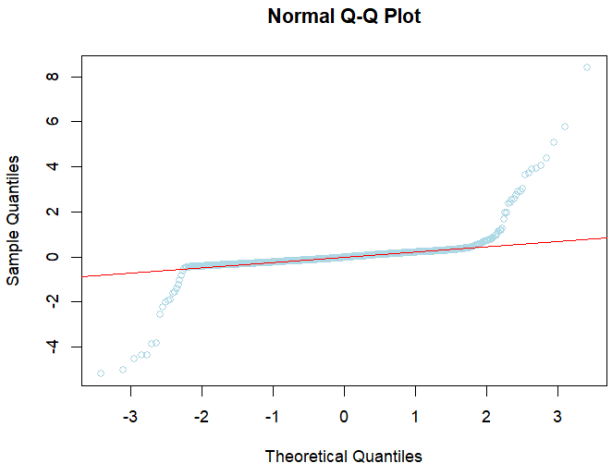


Figure 7:Q-Q plot of Residuals for MLR

Shapiro Wilk Test concluded that the residuals do not follow normal distribution. When we observed the Q-Q plot, it indicates departures from normal distribution assumptions of residuals.

Violation of Normality of residuals assumptions in the MLR model indicate the need for more advanced machine learning techniques to model the sleep time.

Tree Based Methods

Tree based methods are suitable to capture nonlinear relationships between predictors and response variable. We examine that the correlation between predictors and response is not very high that gaining the signals for nonlinear relationships. Also, the tree-based methods robust the outliers. We conducted some tree-based algorithms to get more efficient models to predict the sleep time.

Regression Trees

A Regression Tree is a type of decision tree used to predict continuous outcomes. Unlike Multiple Linear Regression (MLR), which assumes a linear relationship, regression trees work by splitting the data into different regions and making predictions within each region.

After training the model, the following RMSE and R squared values are obtained,

	Training	Testing
RMSE	0.7350	0.7923
R Square	0.6098	0.5436

Table 3:Evaluation metrics for Regression Tree

When assessing the model accuracy, since the train RMSE and test RMSE are closer, the R square value is somewhat low. **Only about 55% of total variation is explained by the model**.

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. This algorithm creates multiple subsets of the training data by randomly sampling. Each subset is used to train a decision tree independently. At each tree split, only a random subset of predictors is considered. The final output is the average of all tree predictions.

We perform random forest model on sleep time data and following results are observed.

	Training	Testing
RMSE	0.2783	0.5810
R Square	0.9522	0.7587

Table 4:Evaluation metrics for Random Forest before eliminating unimportant variables

Training RMSE is relatively low when compared to test MSE. And the R square value for training set is considerably high. So, we can suggest that the model has been overfitted.

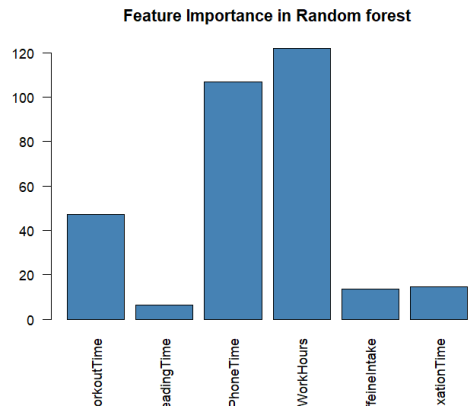


Figure 8: Feature Importance plot in Random Forest

We computed the Feature Importance and got the following plot, Reading Time, Caffeine Intake and Relaxation Time show less importance in the model. We removed these variables and fit the random forest again, then these results are obtained,

	Training set	Test set
RMSE	0.3181	0.6991
R square	0.9321	0.6504

Table 5:Evaluation metrics for Random Forest after eliminating unimportant variables

Results suggested that they still have an overfitted model that did **not show any increment in the accuracy of the model**.

XGBoost

XGBoost (eXtreme Gradient Boosting) is an advanced machine learning algorithm that is faster, more efficient, and more accurate than traditional boosting methods. XGBoost is based on Gradient Boosting, where multiple decision trees are sequentially built to correct the errors of previous trees. So, it prevents overfitting.

Upon training the model and tuning hyperparameters the following results obtained,

	Training set	Test set
RMSE	0.4556	0.5642
R Square	0.8500	0.7685

Table 6: Evaluation metrics for XGBoost

The Test MSE is relatively low when compared with previous models. Also, the overfitting has prevented as training RMSE and test MSE closer. **77% of total variation is explained by the model**.

Support Vector Regression

SVR can handle non-linearly separable data by using kernels to transform input features into a higher-dimensional space where they become linearly separable. It also robust for overfitting. Since the SVR is more sensitive to scale, we first scale the dataset and then apply SVR.

The results we obtained are given below,

	Training set	Test set
RMSE	0.5323	0.5333
R square	0.7952	0.7932

Table 7:Evaluation metrics for SVR

When observed the train RMSE and test RMSE both are close and Test MSE is relatively low when compared to other models. R square value is also significantly high in the model that **approximately 80% of total variation is explained by the model.**

Issues Encountered and Proposed Solutions

During the sleep time prediction analysis, several challenges were identified and addressed to improve model performance and ensure reliability in predictions. Data quality issues, such as inconsistent entries and unusual values, were managed through data filtering and standardization. The Multiple Linear Regression (MLR) model violated the normality assumption of residuals, as confirmed by the Shapiro-Wilk test and Q-Q plot, necessitating the use of more flexible machine learning models like tree-based methods and SVR. Overfitting

was observed in the Random Forest model, with a significantly lower training RMSE (0.2783) compared to the test RMSE (0.5810) and an inflated training R^2 (0.9522). To mitigate this, feature selection was conducted by removing less important predictors (ReadingTime, CaffeineIntake, and RelaxationTime), but overfitting persisted, leading to the exploration of alternative methods such as XGBoost and SVR. Additionally, some predictors showed zero importance in tree-based models, prompting their removal to enhance predictive power. Hyperparameter tuning was crucial for XGBoost, where cross-validation was used to optimize parameters, resulting in a well-balanced model with an improved test R^2 of 0.7685. Sensitivity to scaling in Support Vector Regression (SVR) was another challenge, which was addressed by applying feature scaling before model fitting, leading to the best-performing model with a test RMSE of 0.5333 and a test R^2 of 0.7932. By systematically resolving these challenges, the predictive accuracy of the sleep time model was significantly improved, with the final SVM model emerging as the most robust choice, effectively handling non-linear relationships and maintaining strong generalization performance.

Discussion and Conclusions

The R2 and RMSE values of all types of optimal models (on Test data and Training data) are summarized as follows:

Model	Train		Test	
	RMSE	R square	RMSE	R square
MLR	0.5800	0.7569	0.5523	0.7782
Regression Tree	0.7350	0.6098	0.7923	0.5436
Random Forest	0.3181	0.9321	0.6991	0.6504
XG Boost	0.4556	0.8500	0.5642	0.7685
SVM	0.5323	0.7952	0.5333	0.7932

Table 8:Summary of all R^2 & RMSE values

Model Comparison Plot

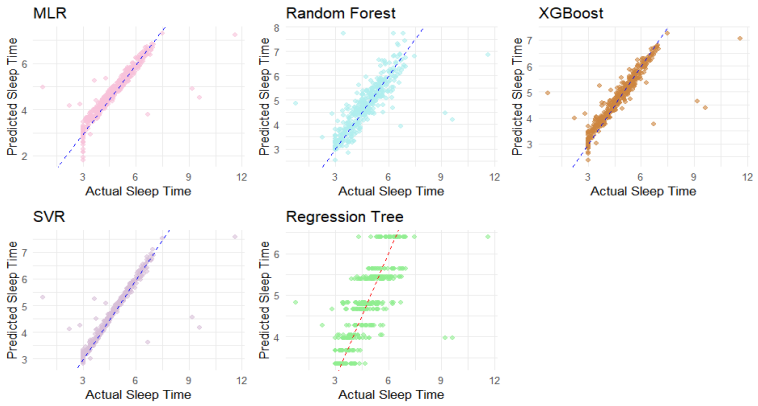


Figure 9:Model Comparison Plot

As uncovered during the Advanced Analysis, SVR is the best model for predicting the Sleep time relatively high-Test $R^2$and lowest testing RMSE among all algorithms, also a minimal difference between train and test R^2 implying that control of over-fitting.

We tuned the hyperparameters and got the optimal SVR model to predict the Sleep Time with highest accuracy.

Model assessing using other evaluation metrics

- **Test MAE: 0.15**
- **Test RMSLE: 0.1**
- **Test MAPE: 4.38%**
- **Test Adjusted R squared: 0.79**

References

Prevalence of Healthy Sleep Duration among Adults — United States, 2014-February 19, 2016 -Yong Liu, MD1; Anne G. Wheaton, PhD1; Daniel P. Chapman, PhD1; Timothy J. Cunningham, ScD1; Hua Lu, MS1; Janet B. Croft, PhD1

Ministry of Health and Welfare; Korea Centers for Disease Control& Prevention. Korea health statistics 2019: Korea national health and nutrition examination survey [KNHANESIV-7]. Sejong: Ministry of Health and Welfare. Korea Centers for Disease Control & Prevention; 2020 [cited 2021 Februay 5].

Available from: https://knhanes.kdca.go.kr/knhanes/sub04/sub04_04_01.do .

Gerber M, Brand S, Herrmann C, Colledge F, Holsboer-Trachsler E, Pühse U. Increased objectively assessed vigorous-intensity exercise is associated with reduced stress, increased mental health and good objective and subjective sleep in young adults. Physiology & Behavior 2014;135:17–24. [doi: 10.1016/j.physbeh.2014.05.047]

Appendix:https://www.kaggle.com/datasets/govindaramsriram/sleep-time-prediction

R Codes: Data-Analysis-Project-2/Advanced analysis project group 14 codes.R at main · pasindugamage2001/Data-Analysis-Project-2