Research on Sleep Health Prediction and Algorithms Based on Big Data

Li Mo[1,*]

[1] High School Affiliated to Shanghai Jiao Tong University, Shanghai, 200439, China
*mo.li.22@jdfzib.org

Abstract. Sleep helps our body recover and wake up full of power. It is also the time when we are growing, both physical and mental. Unfortunately, sleep disorders can prevent individuals from getting adequate rest. 27% of the population in the world have sleep disorders. This can affect their daytime activities. This study aims to predict sleep health using data on daily habits and body conditions and evaluate two algorithms' performance. This study uses Logistic Regression and Random Forest. These two algorithms both perform quite excellent in prediction, so this study tries to determine which one is better and more precise. The results show that Random Forest is more suitable. Not only does Random Forest obtain a higher accuracy score, but it also attains a higher precision score. The Random Forest algorithm achieved a 93% accuracy score in predicting sleep health, with blood pressure being identified as the most important feature in prediction.

Keywords: Sleep Health Prediction, Logistic Regression, Random Forest.

1 Introduction

Sleep health is important for overall health and quality of life. Many people around the world have sleep disorders. It can lead to poor sleep quality and lack of sleep. The most commonly known sleep disorder is insomnia. Other than that, sleep apnea is another harmful sleep disorder. Because sleep helps support basic body functions, and growth, restore energy levels, and other things, people with sleep disorders often encounter difficulties while awake. For example, they are more likely to become angry, have mental illness, and perform tasks poorly.

Furthermore, not getting enough sleep increases the chances of experiencing cardiovascular issues that are undesired. Therefore, sleep health is a huge problem that we are now experiencing. It has a significant impact on our lives.

Studying the prediction of sleep health and algorithms holds immense academic value and should not be underestimated. Some scientists have already started researching this topic, but due to each person's differences, there is no definite conclusion. In this study, machine learning algorithms Logistic Regression and Random Forest were used to predict sleep health. This is very different from the past. Not a lot of research chooses to use both Logistic Regression and Random Forest.

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_7

This research requires extensive knowledge of Python and a significant investment of time and energy. Because each person is a different individual, sleep health prediction requires personalized data and models. In addition, Random Forest and Logistic Regression were never used before in research about sleep health prediction. Sleep disorders are a serious concern as they can lead to loss of productivity and accidents in daily life.

The aim of the study is to predict sleep health based on individuals' daily habits and physical conditions and evaluate the performance of the algorithms. Sleep health was often overlooked in the past and the ways to evaluate it was primate. Each person reacts and sleeps differently. Some might feel comfortable from a short nap, others might sleep better from a whole night's sleep. To improve the accuracy of the results, this study uses machine learning algorithms for modeling, performance comparison, and feature analysis. As a result, the data could be classified before prediction, and thus provide a more accurate and precise result.

This essay is divided into five sections, including "Introduction". The second section, titled "Related Work", focuses on past research related to sleep health. The third section, "Methodology", outlines the research method used. The fourth section, "Experiment and Result", presents the concept and results of sleep health research. The "Conclusion" segment serves as a brief recap of the study and provides insights into prospective applications of the findings.

2 Related Work

There has been extensive research to show the significant meaning of sleep health to humans. With the data collected from students, research shows in great depth some relationships related to sleep quality [1]. Predicting sleep quality through the use of electronic health records and heart rate variability (HRV) techniques is a cost-effective and user-friendly method [2]. Through the analysis of central sleep apnea (CSA) patients, healthy people, and obstructive sleep apnea (OSA) patients, scientists gain more insight into CSA. They can determine the severity and diagnosis of CSA [3]. People with insomnia symptoms and short sleep duration are more likely to experience heart health problems, according to research [4]. Much research has looked into sleep health, both to find out the relation with other factors and to have a better understanding of a specific sleep disorder.

There has been a lot of research conducted to predict sleep health through the use of AI-based methods. Scientists have developed a new algorithm that uses ensemble learning to predict sleeping efficiency on a personalized level. This algorithm can adapt and improve daily by analyzing individual features, but it requires a significant amount of data to function accurately due to the unpredictability of sleep health [5]. Using supervised learning algorithms like Logistic Regression, Random Forest, and Extreme Gradient Boosting, middle-aged and older adults' depressive symptoms can be predicted based on their sleep data. This prediction method is highly accurate [6]. A research used classification algorithms and data collected about students to predict their well-being and resulted in an accuracy of 62%, much lower than this research

[7]. The prediction of sleep apnea can be improved by using both electroencephalogram (EEG) signals and machine learning algorithms. However, this method is only accurate with specific conditions, its accuracy is highly unstable [8]. Scientists have come up with various AI-based methods, both new and old, to predict sleep health. They all have some breakthroughs and shortcomings.

3 Methodology

3.1 Logistic Regression

Logistic regression is a type of supervised machine learning algorithm primarily used for classification tasks. Its purpose is to predict the probability of an instance belonging to a given class [9]. It uses the output of the linear regression as input and uses the function expressed by Equation (1) to estimate the probability of a given class. In Equation (1), the probability of success is represented by p(X), and the variables are denoted by x. The set of coefficients for the independent variables is denoted as β1 [10].

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{1}$$

There are three types of Logistic Regression: binomial, multinomial, and ordinal. Different types of Logistic Regression have different numbers and orders of dependent variables.

Logistic Regression is a valuable decision-making tool that is frequently employed to calculate odds ratios in scenarios involving multiple independent variables. This method determines how each variable impacts the analyzed event's odds ratio [11].

Logistic Regression also has some disadvantages. For example, it can only be used for discrete numbers. And when the relationship involved is too complicated, the algorithm cannot easily operate.

In this study, the dependent variable is “Sleep Disorder”. It is the target that is trying to predict. The rest of the columns on the dataset are the independent variables. They are the input, used to predict the target.

3.2 Random Forest

Random Forest is an ensemble method, creating multiple models and combining them to solve the problem, capable of performing both regression and classification tasks with the use of multiple decision trees and Bootstrap and Aggregation. Bootstrap Aggregation is more commonly known as bagging. It is similar to Random Forest. It computes a predictor from each of the bootstrap samples and then aggregates it into a consensus predictor. The ensemble method and the technique of Bootstrap Aggregation can both help improve the result of the algorithm. As a result, Random Forest is usually a better classifier.

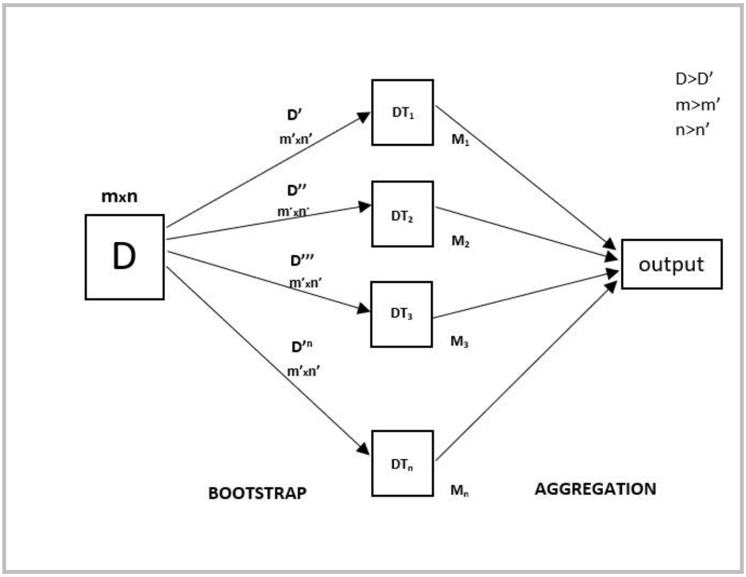

Figure 1: Random Forest Regression Model [12]

As shown in Figure 1 the model for Random Forest. There are multiple decision trees in the regression. Although every decision tree has high variance, when they are combined, the resultant variance will become low. Therefore, the output will not depend on a single decision tree but on multiple decision trees. The first part of the model is called Bootstrap. It predicts each decision tree. When Random Forest is used for classification, the result is the majority voting classifier, when used for regression problems, the final output is the mean of all the output. This is the latter part of the model, called Aggregation.

This feature of Random Forest gives it more randomness. Everything is possible with Random Forest. It doesn't consider the majority, but it is purely random. Random Forest is also much easier to use and more accurate than the decision tree algorithm. It can cope with datasets with missing data, outliers, and noisy features. Its result is often more precise than other algorithms.

It also has some disadvantages, sometimes Random Forest needs a long time to operate because there are too many branches in the model. It is difficult to interpret. Moreover, when there is a large dataset, it is computationally expensive.

3.3 Differences Between the Two Algorithms

First of all, Random Forest can be used for both classification and regression problems, but Logistic Regression is suitable for only binary classification problems. The way these two make predictions is also different. Random Forest is based on multiple decision trees, and Logistic Regression uses the logistic function. Logistic regression is computationally faster than Random Forest. On the matter of money, Random Forest can handle large amounts of data, even if there are missing values, outliers, and non-linear relationships, but Logistic Regression needs more money than Random Forest when there are a large amount of data.

4 Experiment and Result

4.1 Independent and Dependent Variables

The dataset Sleep Health and Lifestyle Dataset used for this research comes from Kaggle.com. The dataset is synthetic. Table 1 shows the names of the columns in the dataset and their explanations [13]. Most columns in the dataset are rated subjectively, such as quality of sleep and stress level.

Table 1. Dataset Columns

Columns	Explanation
Person ID	A number for each participant
Gender	The gender of the person (Male/Female)
Age	The age of the person
Occupation	The occupation of the person
Sleep Duration	The number of hours the person sleeps per day
Quality of Sleep	A subjective rating of the quality of sleep, ranging from 1 to 10
Physical Activity Level	The number of minutes the person engages in physical activity daily
Stress Level	A subjective rating of the stress level experienced by the person, ranging from 1 to 10
BMI Category	The BMI category of the person (Normal, Normal Weight, Overweight, Obese)
Blood Pressure	The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure
Heart Rate	The resting heart rate of the person in beats per minute
Daily Steps	The number of steps the person takes per day
Sleep Disorder	The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea)

The “Sleep Disorder” is the dependent variable, that is the target that is trying to predict. The other columns are the independent variables. In total, there are 400 rows and 13 columns in the dataset.

4.2 Data Preprocessing

In the beginning, a few libraries need to be imported. For example, Pandas, Numpy, Sklearn, Mataplotlib, and Seaborn. Before processing the data, we checked for null or missing values in the dataset. Fortunately, no missing values were found in this dataset.

After the preparation, the dataset needs to be preprocessed. First, columns that are not consistent need to be changed. For instance, the “Blood Pressure” column. The highest blood pressure and lowest were together. In this way, it is difficult to see and classify it. The first step was to split the "Blood Pressure" column into separate ones. The “Blood Pressure” column was dropped, and in its place were “bp_lower” and “bp_upper” for each person. Other than the “Blood Pressure” column, the “Person ID” column was also dropped because it is useless in prediction.

Then, a special function was used for creating dummy variables. A binary variable that indicates if a categorical variable takes on a specific value is called a dummy variable [14]. Dummy variables were created for the categorical columns “Gender”, “Occupation”, and “BMI Category” and they were assigned to the DataFrame “dummies”. Now, before scaling and transforming the input, they were copied.

A class called “CustomScaler” was defined. The columns selected in the class were then scaled and fitted. After that, the mean and variance of the columns were calculated, and they were returned. They were concatenated with the remaining columns of the input data. Using another code, the columns were reordered so that they could later be used for prediction. The order for them is “Age”, “Sleep Duration”, “Quality of Sleep”, “Physical Activity Level”, “Stress Level”, “Heart Rate”, “Daily Steps”, “bp_upper”, “bp_lower”, “Gender_Female”, “Gender_Male”, “Occupation_Accountant”, “Occupation_Doctor”, “Occupation_Engineer”, “Occupation_Lawyer”, “Occupation_Nurse”, “Occupation_Other”, “Occupation_Salesperson”, “Occupation_Teacher”, “BMI Category_Normal”, “BMI Category_Obese”, “BMI Category_Overweight”, “Sleep Disorder”. The next step after reordering is to name them “unscaled”. All rows and columns in the “unscaled” except the last column were selected and assigned as “unscaled_input”. Gender, occupation, and BMI categories were omitted and the rest of the columns in “unscaled_input” were put in the class “CustomScaler” for scaling and transforming.

The column "Sleep Disorder" was designated as the "target". The "scaled_input" and "target" were then inputted into the "train_test_split" function to create training and testing sets. The data was split with 80% used for training and 20% used for testing.

After all this, LogisticRegression and RandomForestClassifier were imported. The training data was fitted into the models. Then, the accuracy of the models was calculated. Using the “predict” method, and putting the test data as input, the target variable, “Sleep Disorder” was predicted.

4.3 Analysis

Before predicting, a series of analyses were done on the data to gain more insight. This helps have a better understanding of the dataset.

Sleep Disorder

For the column “Sleep Disorder”, there are three conditions: None, Sleep Apnea, or Insomnia. Sleep Apnea is a kind of sleep disorder in which an individual suffers from breathing problems during sleep. Insomnia is difficulty falling asleep or staying asleep. As shown in Figure 2 is the distribution of sleep disorder. These three types of conditions are displayed on the x-axis, and the y-axis shows the number. The color blue is for people with no sleep disorder, orange is for people with sleep apnea, and green is for people with insomnia. Figure 2 shows that more than 200 participants are in the “none” category. As for the people with sleep disorders, there is an almost equal number of people with sleep apnea and insomnia. The number is about 70.

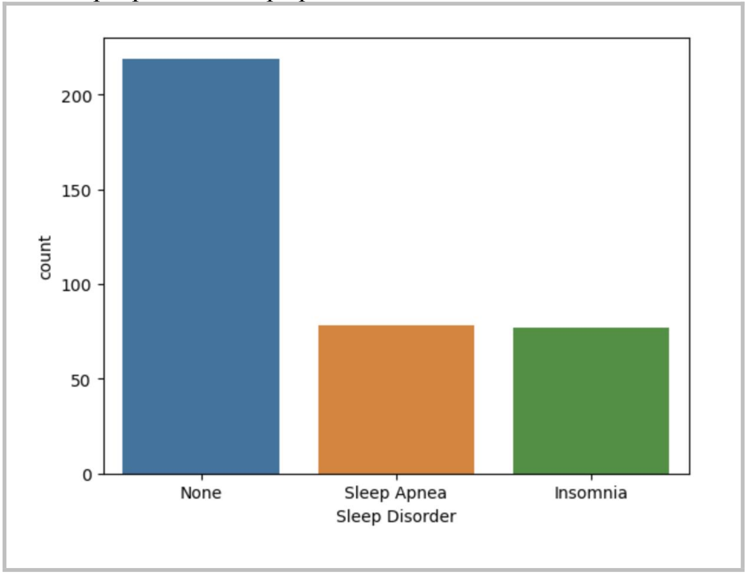

Figure 2: the distribution of sleep disorders (Photo/Picture credit: Original)

Sleep Disorder and Gender

There is a big difference in gender. It is shown that women are 40% more likely to have insomnia than men. As shown in Figure 3, male and female each has three bars, which indicates their own distribution of sleep disorder. Like Figure 1, blue is for no sleep disorder, orange is for sleep apnea, and green is for insomnia. On the left is the distribution for males and the right is for females. The number of males that have no sleep disorder is between 130 and 140, which is almost twice as much as females. In addition, less than 20 males have sleep apnea and around 40 males have insomnia, which is almost the same as females. However, there are more females with sleep apnea, between 60 to 80 people.

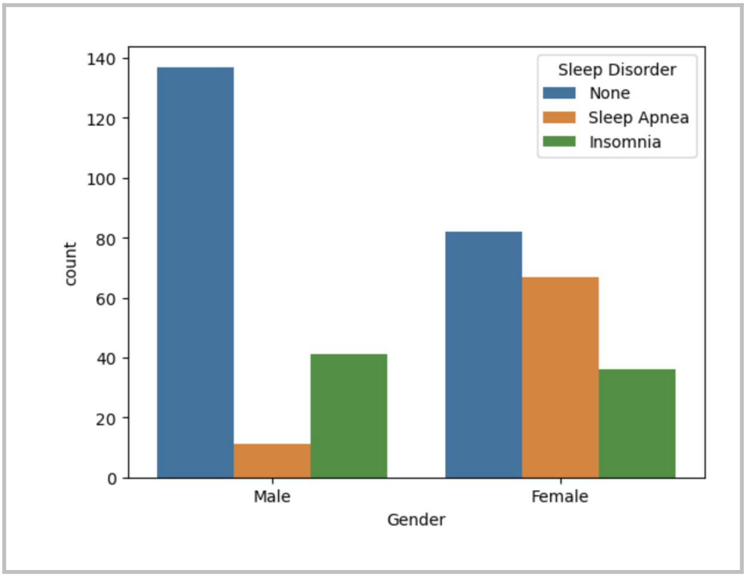

Figure 3: the distribution of sleep disorders by gender (Photo/Picture credit: Original)

Sleep Disorder and BMI Category

Sleep disorders not only can be affected by gender but also by BMI (body mass index) category. There are four parts in the "BMI category" column: overweight, normal, obese, and normal weight. These four categories are written on the x-axis. Again, blue is for people with no sleep disorder, orange is for sleep apnea, and green is for insomnia. Only about 10 people among those with normal body mass have sleep disorders, while over 175 people in this category do not have sleep disorders, as shown in Figure 4. In the overweight category, there is the highest number of people with sleep disorders. Both sleep apnea and insomnia have a number between 50 and 75. But, there are also about 25 people who have no sleep disorders. Obese people have the highest possibility of having sleep disorders because although there is only a small number of people in this category, they all have sleep disorders, half with sleep apnea and half with insomnia. For normal weight, it has the same condition as normal: most people have no sleep disorder and only a few people have sleep disorders.

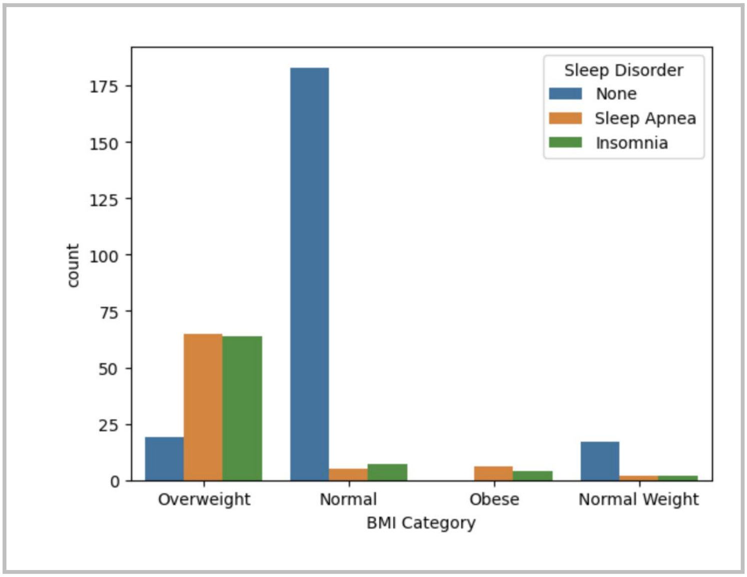

Figure 4: the distribution of sleep disorders by BMI category (Photo/Picture credit: Original)

4.4 Result

The Performance Comparison

Table 3 displays the Logistic Regression classification report, while Table 4 displays the Random Forest classification report. The two tables contain the results obtained by the two models. As shown in Table 3 and Table 4, The accuracy score and f1-score of Logistic Regression are both 92%. and the accuracy score and f-1 score obtained by Random Forest is 93%, 1% higher than Logistic Regression. The scores achieved by the two models are both very good, with Random Forest slightly higher. The precision for Logistic Regression is 92%, and for Random Forest is 94%.

Table 3. Classification Report of Logistic Regression

	precision	recall	f1-score	support
Insomnia	0.93	0.88	0.90	16
None	0.92	0.96	0.94	48
Sleep Apnea	0.90	0.82	0.86	11
accuracy			0.92	75
macro avg	0.92	0.88	0.90	75
weighted avg	0.92	0.92	0.92	75

Table 4. Classification Report of Random Forest

	precision	recall	f1-score	support
Insomnia	1.00	0.88	0.93	16
None	0.92	0.98	0.95	48

Sleep Apnea	0.90	0.82	0.86	11
accuracy			0.93	75
macro avg	0.94	0.89	0.91	75
weighted avg	0.94	0.93	0.93	75

Feature Importance

The feature importance of sleep health is shown in Figure 4. Based on the numbers shown on the x-axis, the importance of a specific feature is shown. The longer the bar, the more important the feature. The y-axis displays the name of the input. Every bar has a different color and length. Blood pressure is the most important. This feature has an importance level greater than 0.175. After blood pressure, the next most important factor is the BMI category, with a feature importance of almost 0.175. In third place is occupation, with a feature importance of around 0.15. Age and sleep duration have similar feature importance, with age having a slightly higher importance. All the others' feature importance is less than 0.1. The last on the list is gender, its feature importance is less than 0.025.

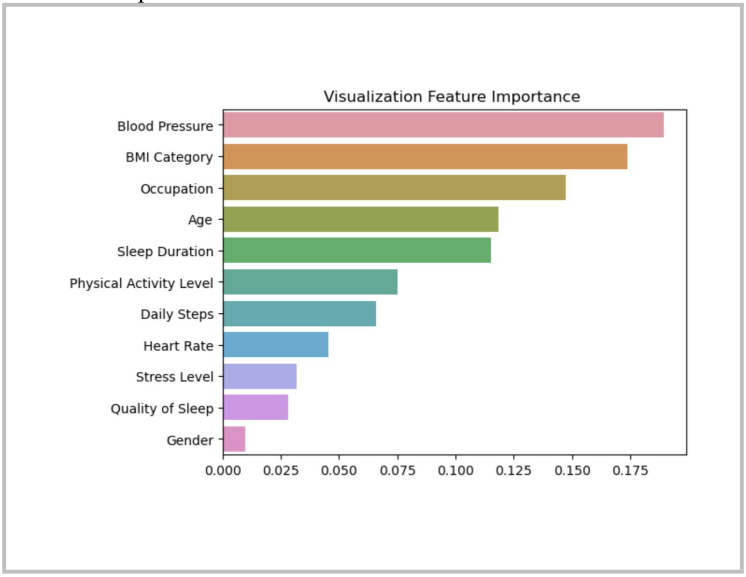

Figure 5: the feature importance of sleep health [15]

5 Conclusion

Currently, sleep health has caught the eyes of people from various fields. Sleep health is related to both our physical health and mental health. People with sleep disorders often have low quality or difficulty falling asleep. This results in poor behavior during the daytime. Lots of research has been trying to discover the specific relationship between sleep health and other aspects of human health. Much has been done to have a deeper understanding of sleep health.

However, sleep health is not that easy to predict. Everyone has a different body condition. Some rules can apply to one group of people but not another. Therefore, there is a great need for an accurate and precise way of prediction.

This research utilized two machine learning algorithms for predicting sleep health with greater accuracy and precision. The Logistic Regression and Random Forest algorithms are commonly used for classification. Multiple variables were used for the prediction, with the target variable being "sleep disorder", which includes three conditions: individuals without sleep disorders, those with sleep apnea, and those with insomnia. The study shows that Random Forest is more suitable for sleep health prediction. It has an accuracy score of 93% and a precision score of 94%. Logistic Regression achieved a 92% accuracy score and precision in predicting sleep health. For both models, the accuracy score and f-1 score are the same, that is because the data was fitted during preprocessing.

Other than the performance of the two models, the most important feature in prediction has been shown to be blood pressure. For future research and prediction, blood pressure is a data that is crucial. Body mass is another important feature. To prevent sleep disorders, the best way is to exercise regularly, have a healthy diet, and regulate your sleeping routine.

This study has significant meaning. Using machine learning algorithms to predict sleep health is a big breakthrough. Finding out that Random Forest is the more suitable model for this prediction is another. The Random Forest method provides unmatched precision and is frequently employed for identifying feature importance. Nevertheless, this study has some limitations. For example, this study only uses the data collected from 374 participants.

The methods and models used in this research can be applied by others to gain more insight into predicting sleep health and sleep health itself. Based on this research, future research could try to find ways to solve or reduce sleep disorders. Hopefully, there will be more knowledge about sleep health, and people can have better bodies.

References

1. Sharisha Shanbhog, M., Jeevan Medikonda.: A clinical and technical methodological review on stress detection and sleep quality prediction in an academic environment. Computer Methods and Programs in Biomedicine 235, 107521 (2023).
2. Sadeghi, R., Banerjee, T., Hughes, J.: Predicting Sleep Quality in Osteoporosis Patients Using Electronic Health Records and Heart Rate Variability. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 5571-5574 (2020).
3. Dvir H. et al.: Central Sleep Apnea Alters Neuronal Excitability and Increases the Randomness in Sleep-Wake Transition. in IEEE Transactions on Biomedical Engineering 67 (11), 3185-3194 (2020).
4. Daniel J. Biddle, Daniel F. Hermens, Tea Lallukka, Melissa Aji, Nick Glozier.: Insomnia symptoms and short sleep duration predict trajectory of mental health symptoms. Sleep Medicine, 53-61 (2019).

5. Van, N. T. P., Son, D. M., Zettsu, K.: A Personalized Adaptive Algorithm for Sleep Quality Prediction using Physiological and Environmental Sensing Data. 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), 113-119 (2021).
6. Stephania Ruth Basilio Silva Gomes, Malcolm von Schantz, Mario Leocadio-Miguel.: Predicting depressive symptoms in middle-aged and elderly adults using sleep data and clinical health markers: A machine learning approach. Sleep Medicine, 123-131 (2023).
7. Akif Can Kılıç, Ahmet Karakuş, Emre Alptekin.: Prediction of University Students' Subjective Well-Being with Sleep and Physical Activity Data using Classification Algorithms. Procedia Computer Science, 2648-2657 (2022)
8. Onargan, A., Gavcar, B., Çalışkan, G., Akan, A.: Prediction of Sleep Apnea Using EEG Signals and Machine Learning Algorithms. 2021 Medical Technologies Congress (TIPTEKNO), 1-4 (2021).
9. GeeksforGeeks, https://www.geeksforgeeks.org/understanding-logistic-regression/, last accessed 2023/07/14.
10. Kairo Pereira Teodoro da Silva, Andreza Kalbusch, Elisa Henning.: Detection of unauthorized consumption in water supply systems: A case study using logistic regression. Utilities Policy 84, 101647 (2023)
11. Sperandei, S.: Understanding logistic regression analysis. Biochem. Med 24(1), 12-18 (2014)
12. GeeksforGeekd, https://www.geeksforgeeks.org/random-forest-regression-in-python/, last accessed 2023/06/05.
13. Kaggle,https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset?resource=download, last accessed 2023/05/23.
14. GeeksforGeeks, https://www.geeksforgeeks.org/how-to-create-dummy-variables-in-python-with-pandas/#:~:text=A%20dummy%20variable%20is%20a%20binary%20variable%20that,dummy%20variables%20in%20python%20using%20get_dummies%20%28%29%20method., last accessed 2022/01/16.
15. Kaggle, https://www.kaggle.com/code/raullloriz/sleep-health-eda-predict-random-forest, last accessed 2023/08/06.