Data Science Seminar (DASC6810) Project:

Sleep Cycle and Productivity Analysis using Machine Learning

Jiayi Zhang (T00752757)
Shashank Manjunatha (T00728166)

February 15th, 2024

Table of Contents

Abstract

Sleep patterns play a significant role in deciding one's productivity. This study employs machine learning techniques to analyze the impact of sleep cycles and lifestyle on productivity scores. From a well-structured dataset, different techniques like Decision Trees, Random Forest, and Principal Component Analysis (PCA) are employed to find out the influential factors. The results highlight the salient variables determining productivity and offer a best-fit predictive model. From inspection of sleep patterns, quality of sleep, workout routine, levels of caffeine consumption, and nighttime screen time, we demonstrate the association between measures and productivity rates.

This study is motivated by the increasing worry about loss of productivity because of sleep disorders, particularly in jobs that involve heavy cognitive effort. The data used are a variety of features that range from sleep start and end times to sleep quality ratings, lifestyle choices like caffeine consumption and exercise, and productivity metrics like work duration and self-assessed efficiency.

Three unrelated machine learning models were employed to calculate the relationship between sleeping patterns and productivity. Decision Tree Regression provides an interpretable method of understanding the impact of each variable, while Random Forest Regression provides more predictive power through overfitting suppression by ensemble learning. Principal Component Analysis (PCA) supports feature selection through the identification of the most influencing factors on the variation in productivity. Research indicates that the top predictive model is Random Forest Regression. Findings are useful in helping organizations and professionals schedule work efficiently and better the wellbeing of employees through facilitating improved sleep behaviors.

Introduction

Sleep is a crucial aspect of human health and functioning, with direct effects on mental function, mood, and productivity. The modern lifestyle, which is characterized by excessive screen exposure, unregulated working hours, and immense stress, has spawned widespread sleep disorders. Scientific studies have consistently demonstrated that sleep loss leads to memory lapses, decreased concentration, and decreased problem-solving ability. Hence, there is a need to understand the intricate connection between sleep patterns and productivity in a bid to enhance workplace effectiveness and individual performance.

Productivity is a key indicator of success at the workplace and in school. Productivity relies on a series of determinants, and quality of sleep, work schedule, degree of stress, and other daily routines such as coffee drinking and exercise are among them. There is evidence to suggest that individuals who maintain normal sleeping habits have better cognitive capabilities compared to inconsistent sleeping patterns. Additionally, poor sleep may cause long-term tiredness, increased disease risk, and long-term psychological issues such as depression and anxiety.

Advances in data-driven technologies have enabled scientists to utilize machine learning models to predict behavior trends and make forecasts with a high degree of accuracy. From a robust dataset that contains unique sleep and lifestyle traits, the paper aims to determine how various parameters determine productivity levels. The information includes crucial variables such as total sleep duration, sleep quality scores, amounts of caffeine intake, screen exposure before going to bed, and occupational characteristics such as working hours and stress. Correlations of sleep cycles with productivity can be ascertained by analyzing the information.

This study employs a set of Machine Learning algorithms, including Decision Tree Regression, Random Forest Regression, and Principal Component Analysis (PCA), to analyze the influence of sleep and lifestyle factors on productivity. Decision Trees provide the intuitive ease of interpreting variable importance, while Random Forest enhances the predictive accuracy by employing an ensemble approach. PCA, in turn, facilitates variable selection through identifying the major factors contributing to productivity variation. By comparing their performances, we aspire to identify the most suitable approach for the prediction of productivity from sleep data.

The primary objective of this research is to provide actionable suggestions for professionals, employers, and individuals who seek to optimize their sleeping habits for improved efficiency. Organizations can utilize the results to implement better work schedules, improve the well-being of their employees, and create a stronger workforce. In addition, individuals can make educated lifestyle choices to improve their cognitive performance and overall productivity.
Overall, this study seeks to bridge the divide between machine learning application and sleep science by analyzing structured data to identify sleep and productivity trends and relationships. The outcomes contribute to the general literature on maximizing workplace productivity and

offer practical recommendations for improved sleep hygiene. The predictive models may be further enhanced by future research by applying sophisticated deep learning techniques and physiological measures of sleep such as EEG recordings.

As the world grows more complex, it is increasingly vital to study the relationship between sleep patterns and productivity. Employees are struggling to maintain work and personal life balance, which causes disrupted sleep cycles and lower production. Companies can leverage data-based methods to personalize solutions in order to enable workers to tailor their sleep schedule to improve overall efficiency and job satisfaction. Moreover, promoting an awareness of the value of sleep can lead to a cultural shift in which productivity is optimized not through longer working hours, but through more intelligent, well-rested workforces.

This research also identifies the potential for integrating wearable technology and real-time monitoring in future studies. Smartwatches and sleep monitors can provide more accurate and continuous data, enabling deeper insights into sleep patterns and how they influence productivity. The application of machine learning to real-time monitoring can result in advanced predictive models that offer personalized sleep recommendations, ultimately improving people's working and personal lives.

Related Work

[1] H. Phan et al., "L-SeqSleepNet: Whole-cycle Long Sequence Modelling for Automatic Sleep Staging," arXiv, vol. 2301.03441, Jan. 2023.

This paper introduces L-SeqSleepNet, a deep learning model designed to improve automatic sleep staging by using long sequence data. Unlike traditional methods that focus on short segments, L-SeqSleepNet models entire sleep cycles, capturing long-term dependencies in sleep patterns. The model outperforms existing techniques in both accuracy and robustness, showcasing its ability to provide more precise sleep stage predictions. The approach has significant potential for integration into wearable devices and home sleep monitoring applications, offering a better understanding of sleep dynamics for health management.

[2] O. Kılıç et al., "Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning," arXiv, vol. 2303.06028, Mar. 2023.

In this study, the authors propose a machine learning method that combines convolutional neural networks (CNNs) and ensemble learning for predicting sleep quality using data from wearable devices. By analyzing accelerometer and heart rate data, their model accurately predicts sleep quality, surpassing traditional techniques. The system is capable of processing real-time data, making it ideal for use in consumer-grade wearables. This work highlights the potential of machine learning to personalize sleep health management and provide tailored recommendations for improving sleep quality.

[3] T. U. Wara et al., "A Systematic Review and Meta-Analysis on Sleep Stage Classification and Sleep Disorder Detection Using Artificial Intelligence," arXiv, vol. 2405.11008, May 2024.

This paper provides a comprehensive review and meta-analysis of AI-based methods for sleep stage classification and sleep disorder detection. It evaluates various machine learning algorithms, such as deep learning and support vector machines, to assess their performance in detecting sleep stages and disorders like sleep apnea, insomnia, and narcolepsy. The study highlights the significant improvements AI brings to sleep medicine, enabling more accurate diagnoses. It also emphasizes the need for large, high-quality datasets and the potential benefits of integrating multimodal data to further improve AI models.

[4] D. A. Almeida et al., "A Machine-Learning Sleep-Wake Classification Model Using a Reduced Number of Features Derived from Photoplethysmography and Activity Signals," arXiv, vol. 2308.05759, Aug. 2023.

The authors propose a machine learning model that classifies sleep-wake states using a reduced set of features extracted from photoplethysmography (PPG) and activity signals. By simplifying the feature set, the model is computationally efficient, making it ideal for use in wearable devices. Despite using fewer features, the model achieves high classification accuracy, demonstrating its suitability for real-time sleep monitoring applications. The research suggests that such an approach could be widely implemented in consumer devices, offering a practical solution for sleep tracking and health monitoring.

[5] "A Machine-Learning Model for Predicting Sleep and Wakefulness Based on Data from Wearable Sensors," Neuropsychiatric Disease and Treatment, vol. 15, pp. 2943-2952, 2019.
This paper presents a machine learning model that predicts sleep and wakefulness states using wearable sensor data, particularly from accelerometers and heart rate monitors. The authors demonstrate that the model provides highly accurate predictions of sleep stages compared to traditional methods. Their approach shows promise for real-time, continuous monitoring of sleep, with potential applications for early detection of sleep disorders. The study highlights the role of wearable technology in enhancing personalized sleep management and providing actionable insights into sleep health.

[6] "Sleep Better, Live Better: Machine-Learning Method Can Predict Your Body Clock," Neuroscience News, Feb. 2024.
This article discusses a machine learning method designed to predict individuals' circadian rhythms, or "body clock," based on data from wearable devices. By analyzing physical activity, light exposure, and sleep patterns, the model predicts the optimal times for work and rest, helping individuals adjust their daily routines. The research emphasizes the importance of aligning daily schedules with the body's natural rhythms to improve sleep quality and overall health. The approach offers personalized sleep recommendations, with potential applications in optimizing mental and physical well-being through tailored lifestyle adjustments.

[7] "Artificial Intelligence in Sleep Medicine: The Dawn of a New Era," Neuropsychiatric Disease and Treatment, vol. 15, pp. 1777-1783, 2019.
This paper explores the transformative role of artificial intelligence (AI) in sleep medicine, particularly in diagnosing and treating sleep disorders like sleep apnea, insomnia, and narcolepsy. The authors review how AI enhances the analysis of polysomnography (PSG) data, improving the accuracy of sleep stage classification. AI's ability to automate sleep scoring and detect anomalies in sleep data is discussed, along with its potential for use in home sleep testing. The paper argues that AI will play a pivotal role in making sleep health management more accessible and efficient, though challenges such as data privacy and ethics must be addressed.

[8] "Prediction of Good Sleep with Physical Activity and Light Exposure: A Preliminary Study," Journal of Clinical Sleep Medicine, vol. 14, no. 12, pp. 2161-2166, Dec. 2024.
This study investigates the relationship between physical activity, light exposure, and sleep quality using machine learning to predict good sleep outcomes. The authors find that physical activity during the day and appropriate light exposure, particularly in the evening, significantly improve sleep quality. By integrating data from activity trackers and light sensors, the model predicts sleep outcomes with high accuracy. The findings suggest that wearable devices could offer personalized recommendations to optimize sleep health, making it possible to improve sleep patterns through lifestyle changes based on data-driven insights.

Approaches

Decision Tree Regression:

Decision Tree Regression is a supervised learning algorithm that uses a tree-like structure to make predictions. The data is recursively split into subsets based on the most important features, and the final prediction is made by following the decision rules created at each node. It is commonly used for problems where interpretability is crucial.

Features:

- Recursively splits data based on feature importance.
- Makes predictions through decision rules derived from data.
- Handles both categorical and numerical data.
- Captures non-linear relationships between features and target variables.
- Provides high interpretability with human-readable rules.
- Prone to overfitting; requires techniques like pruning to improve generalization.
- Models' complex interactions between features.

Random Forest Regression:

Random Forest Regression is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy. It reduces overfitting by training on random subsets of data, and it provides more reliable predictions by combining the outputs of several trees.

Features:

- Builds multiple decision trees in parallel using bootstrapped samples of data.
- Aggregates predictions from all individual trees to improve accuracy.
- Handles missing data effectively and can still make reliable predictions.
- Performs feature importance analysis to identify the most influential factors.
- Reduces overfitting compared to a single decision tree by averaging results from multiple trees.
- Requires more computational power and time due to training multiple trees.

Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the data into a smaller number of uncorrelated variables, called principal components, while retaining the most significant variance in the data. PCA helps identify the most important features and reduce the complexity of models.

Features:

- Reduces data dimensionality by transforming features into principal components.
- Captures the maximum variance in the data, focusing on the most important variables.
- Standardizes the dataset to ensure equal contribution from all features.
- Removes redundant features, improving model efficiency.
- Addresses multicollinearity by eliminating highly correlated variables.
- Improves computational efficiency by reducing the number of features.

Methodology

1. Data Preprocessing

The first step in the analysis is to prepare the dataset so that it is clean and ready for modeling. This includes several important processes:

Data Cleaning: Irrelevant columns are removed since they are not useful for the analysis. This step removes unnecessary noise from the dataset and makes it computationally efficient.

Encoding Categorical Variables: There are a few categorical variables in the data, such as Gender, that need to be converted into numbers so that they can be recognized by machine learning algorithms. The categorical values are encoded using Label Encoder so that they can be processed by algorithms.

Dealing with Missing Values: Missing data is handled with the proper imputation methods. Numerical values have missing entries filled with the mean or median, and categorical values are filled in with the most frequent category. This completes and makes the dataset strong enough for the modeling.

Identifying Target and Features: The data is divided into independent variables (features) like Age, Total Sleep Hours, Sleep Quality, Exercise, Caffeine Intake, Screen Time Before Bed, and Work Hours. The dependent variable (target) is the Productivity Score, which we are going to predict.

Splitting Data: For efficient training and testing of the machine learning models, the dataset is split into the training and test sets in the ratio 80:20. This gives the model adequate data for training while also reserving some for testing to avoid overfitting and ensure that the model can generalize to new data.

2. Transformation of Features

Following the preprocessing of the dataset, several transformations are performed to prepare the data for optimal model performance:

Feature Scaling: Features in the dataset, such as Caffeine Intake and Screen Time Before Bed, may be on different scales. To avoid bias in the model towards those variables that have larger ranges, StandardScaler is applied to normalize these features so that all of them contribute equally to the learning process of the model.

Dimensionality Reduction: High-dimension data will introduce redundancy and inefficiency. For this problem, Principal Component Analysis (PCA) is employed. PCA eliminates the dimension of the dataset using principal components to capture maximum variance explaining most about the data, thus efficient computation and best model performance.

Feature Engineering: New features are derived from the original data to enhance the model's predictive capability. For example, Sleep Efficiency is calculated by dividing Total Sleep Hours by Work Hours, providing a measure of how well a person is making use of their sleep time relative to work obligations.

3. Model Selection

After preparing the data, different machine learning models are employed to establish the best possible means of studying the relationship between productivity and sleep cycles:

Decision Tree Regression: Decision trees are tree-based models that split the data based on different criteria. Decision trees are best utilized to understand how differently variables contribute towards productivity and assist in offering interpretability by visualization of the decision-making process.

Random Forest Regression: A sophisticated ensemble technique, Random Forest builds numerous decisions trees and makes predictions from them together. This inhibits overfitting and enhances the accuracy of the model by utilizing the experience of numerous trees rather than relying on a single tree.

Principal Component Analysis (PCA): While PCA is primarily thought of as a dimension reduction technique, PCA is also looked at for establishing the alignment of the principal components to productivity. By analyzing the variance explained by these components, we find out which of the components are productivity drivers.

4. Training of models

Having decided upon the models, now it is a matter of training these models on the dataset prepared:

Decision Tree Model: A Decision Tree Regressor is instantiated with a maximum depth of 5 to avoid excessive complexity and overfitting. The model is trained to learn significant patterns in the data.

Random Forest Model: The Random Forest Regressor is instantiated with 100 trees, utilizing ensemble learning to improve predictive accuracy by averaging predictions from multiple decision trees.

PCA Transformation: The data is converted through PCA to find the principal components. They are then analyzed to understand how they correlate with productivity and to find the most important factors.

5. Model Evaluation

Model performance is measured using a range of evaluation techniques:

Decision Tree Analysis: The decision tree model is visualized in order to achieve an intuitive understanding of the decision process. Decision rules are inferred in order to determine how each feature contributes towards productivity prediction. Feature importance scores are also determined to determine the most significant variables.

Random Forest Analysis: Similar to the decision tree, feature importance scores are examined to estimate the contribution of various predictors such as sleep and working hours. Model performance is also evaluated using statistical metrics such as R-squared (R^2) or some other evaluation methods that provide insights into the model's predictive power.

PCA Analysis: The loading of variables onto the principal components is visualized through a heatmap to discern how each variable loads on the components, thereby aiding in the interpretation of the main factors influencing productivity. The correlation analysis identifies the predictors that are strongest, according to the PCA findings.

Results

After running the code of three methods-Decision Trees, Random Forest Regression, and PCA, here are the result of data output and visualization graphs:

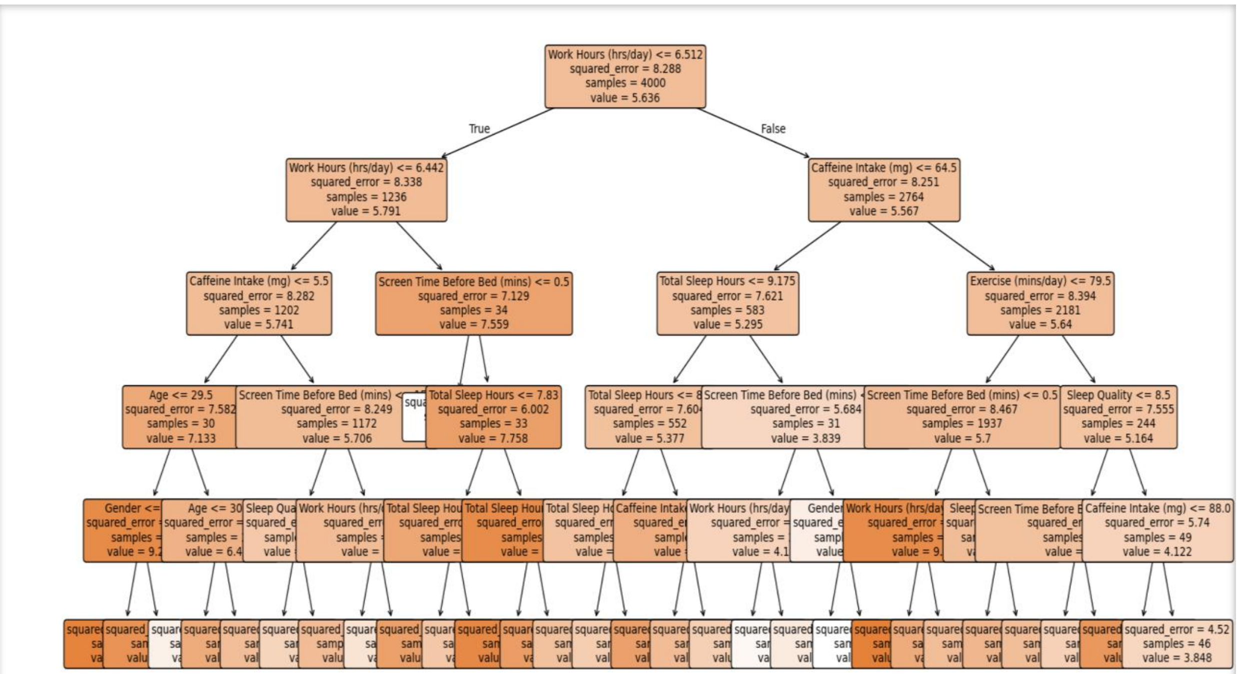
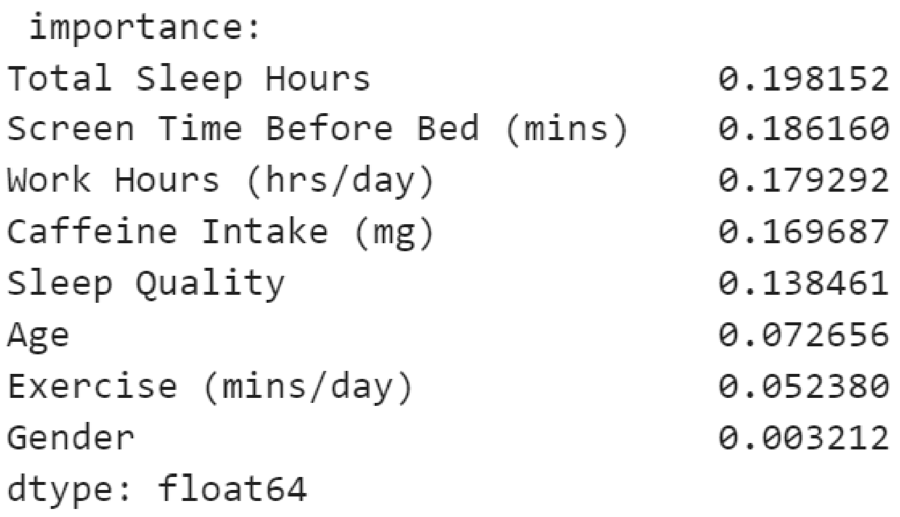


Fig.1. Decision Tree Rules



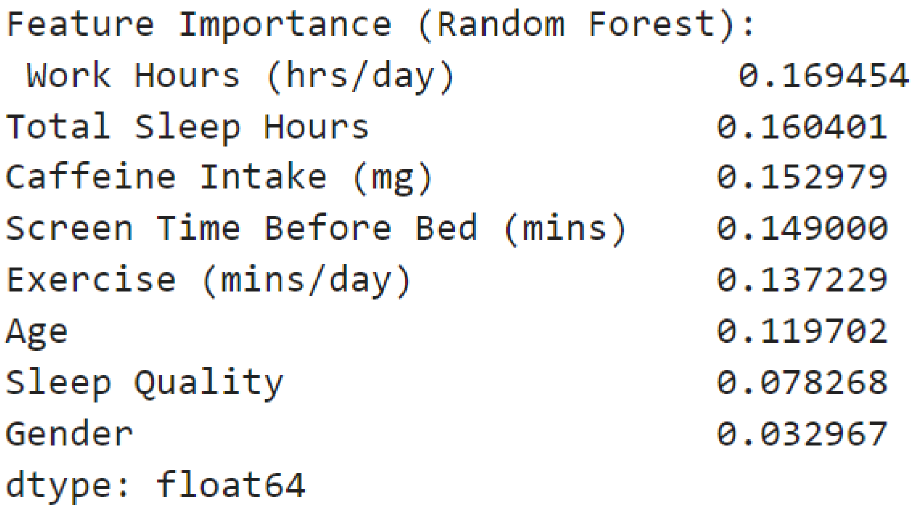
```
 importance:
Total Sleep Hours                0.198152
Screen Time Before Bed (mins)    0.186160
Work Hours (hrs/day)             0.179292
Caffeine Intake (mg)             0.169687
Sleep Quality                    0.138461
Age                              0.072656
Exercise (mins/day)              0.052380
Gender                           0.003212
dtype: float64
```

Fig.2. Result in Decision Tree

From the tree rules graph above, we can apparently see that this decision tree splits data based on different conditions at each node to predict a target variable. From the top we can see the tree root node represents the first split, based on "Work Hours <= 6.512".and each subsequent node represents a decision based on another variable, such as "Caffeine Intake (mg)", "Screen

Time Before Bed (mins)", "Total Sleep Hours", etc. Noticed that the values within each node indicate the squared error, the number of samples in that category, and the predicted value. As the deeper the tree goes, the more specific the predictions become. The color intensity reflects the values, with darker shades representing higher values.

From the form below, we get to acknowledge the rank among these factors by scandalized score values. The total sleep hours are the most crucial factors which scored 0.198, following by screen time before bed and working hours per day. The Gender is the most irrelevant factors which only get 0.003 scores in value. The result roughly meets the reality situation since sleeping properly can intensify the connection between neuron cells and lift productivity and other factors like gender can not influent productivity much because of the evaluation baseline between genders productivity are always shifting. Hence, it provides moderate robust model in this dataset.



```
Feature Importance (Random Forest):
 Work Hours (hrs/day)             0.169454
Total Sleep Hours                0.160401
Caffeine Intake (mg)             0.152979
Screen Time Before Bed (mins)    0.149000
Exercise (mins/day)              0.137229
Age                              0.119702
Sleep Quality                    0.078268
Gender                           0.032967
dtype: float64
```

Fig.3. Result in Random Forest

Compared to decision trees, Random Forest Regression model give us a slightly different ranking among eight factors. The working hours is the most important method in this model. while age, sleep quality, or gender remains low rank in the evaluation. The difference among models will dig in in detail in the conclusion part.

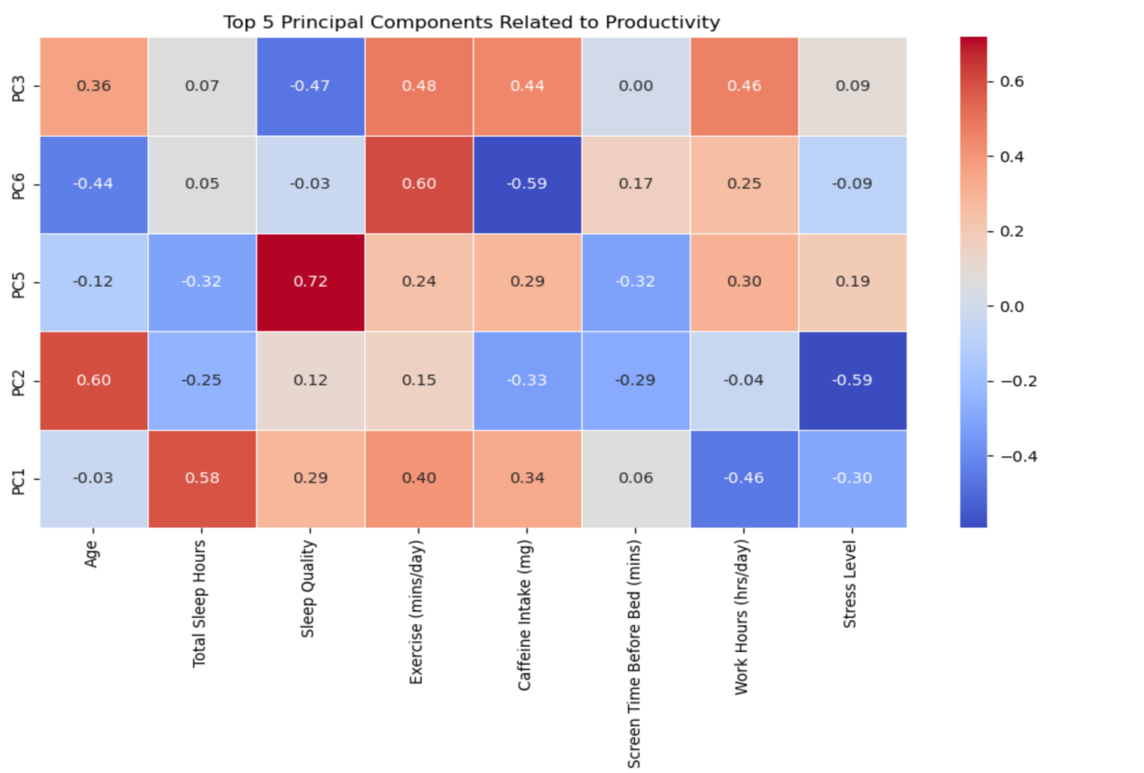
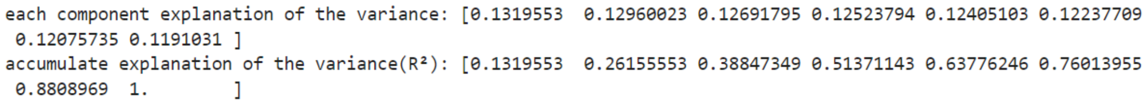


Fig.4. Result in PCA

Taking every most crucial factors in the top 5 principal components. For PC1 it is the total sleep hours, following by Age, working hours, sleep quality, and exercise, in PC2, PC3, PC5, PC6 (These are all the names we made for combined factors, not the actual column names). We believe these 5 factors take up the most explanation of the variance which means it is appropriate to neglect other trivial factors in this dataset.



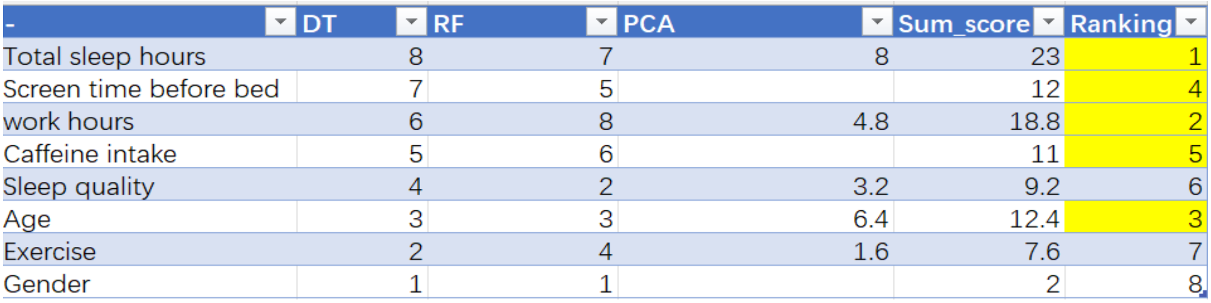
```
each component explanation of the variance: [0.1319553  0.12960023 0.12691795 0.12523794 0.12405103 0.12237709
 0.12075735 0.1191031 ]
accumulate explanation of the variance(R²): [0.1319553  0.26155553 0.38847349 0.51371143 0.63776246 0.76013955
 0.8808969  1.        ]
```

Fig.5. Result for explanation of the variance

And From Fig.5. we can figure out the top- five factors could make 63% of the accumulative variance explanation, which is a satisfying result.

Conclusion

Using the ensemble learning strategy, we can combine the result of the three models together to give an overall instructive conclusion by scoring the rank for each model. The result is below:



-	DT	RF	PCA	Sum_score	Ranking
Total sleep hours	8	7	8	23	1
Screen time before bed	7	5		12	4
work hours	6	8	4.8	18.8	2
Caffeine intake	5	6		11	5
Sleep quality	4	2	3.2	9.2	6
Age	3	3	6.4	12.4	3
Exercise	2	4	1.6	7.6	7
Gender	1	1		2	8

Fig.6.Result of Ensemble score

Because there are 8 factors in the dataset, we manually rank the greatest relevance as 8 and the smallest as 1. For PCA we only pick 5 factors, to equalize the score evaluation, we gave 8,6.4,4.8,3.2,1.6 separately to the 5 factors, which like the process of standardization. After scoring, (The maximum score is 24 and the minimum is 2) we are able to use the functions in the Excel to summarize and calculate the ranking. Among the top-five factors, age is the uncontrollable factors, other four factors can be adjusted by our daily routine changing. With the guidance of the result, we can give the following suggestions:

- Sleep properly every day since Sleep hours is the most relevant factor to the productivity.
- Working in a long time will help you enter heart flow mode which boost productivity.
- Control the screen time and the caffeine intake will help you get better sleep and then boost productivity.

Drawbacks:

i. The dataset contains several additional factors which need further research on the relevance such as the mood score and stress score.

ii. The robust ability of the model needs to be tested by changing different datasets in similar topics.

iii. Several machine learning techniques could be taken in further studies such as XGBoost Forest and Neural Network.

References

[1] H. Phan, K. P. Lorenzen, E. Heremans, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, M. Baumert, K. Mikkelsen, and M. De Vos, *"L-SeqSleepNet: Whole-cycle Long Sequence Modelling for Automatic Sleep Staging," arXiv*, vol. 2301.03441, Jan. 2023.

[2] O. Kılıç, B. Saylam, and Ö. D. İncel, *"Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning," arXiv*, vol. 2303.06028, Mar. 2023.

[3] T. U. Wara, A. H. Fahad, A. S. Das, and M. M. H. Shawon, *"A Systematic Review and Meta-Analysis on Sleep Stage Classification and Sleep Disorder Detection Using Artificial Intelligence," arXiv*, vol. 2405.11008, May 2024.

[4] D. A. Almeida, F. M. Dias, M. A. F. Toledo, D. A. C. Cardenas, F. A. C. Oliveira, E. Ribeiro, J. E. Krieger, and M. A. Gutierrez, *"A Machine-Learning Sleep-Wake Classification Model Using a Reduced Number of Features Derived from Photoplethysmography and Activity Signals," arXiv*, vol. 2308.05759, Aug. 2023.

[5] *"A Machine-Learning Model for Predicting Sleep and Wakefulness Based on Data from Wearable Sensors," Neuropsychiatric Disease and Treatment*, vol. 15, pp. 2943-2952, 2019.

[6] *"Sleep Better, Live Better: Machine-Learning Method Can Predict Your Body Clock," Neuroscience News*, Feb. 2024.

[7] *"Artificial Intelligence in Sleep Medicine: The Dawn of a New Era," Neuropsychiatric Disease and Treatment*, vol. 15, pp. 1777-1783, 2019.

[8] *"Prediction of Good Sleep with Physical Activity and Light Exposure: A Preliminary Study," Journal of Clinical Sleep Medicine*, vol. 14, no. 12, pp. 2161-2166, Dec. 2024.

[9] *Citation: [1] Google, "Google Search," [Online]. Available: https://www.google.com. [Accessed: Feb 2025].*

[10] OpenAI, "ChatGPT, an AI language model," OpenAI, [Online]. Available: https://chat.openai.com/. [Accessed: Feb 2025].

GitHub Repository Link: *https://github.com/ShashankManjunath-717/Sleep-Cycle-and-Productivity-Analysis-using-Machine-Learning_DASC_6810.git*