HDSC WINTER ‘23 PREMIERE PROJECT PRESENTATION: PREDICTING SLEEP EFFICIENCY FOR INDIVIDUALS.

A PROJECT BY TEAM KNN.

OUTLINE.

* Introduction
* Aims & Objectives
* Project Flow
* Conclusion

INTRODUCTION.

Take a moment to consider how you feel when you finally get some sleep after a stressful day. I’m sure there’s a big smile on your face.

Sleep is one of the most fundamental components of everyday life. While the list of its benefits has not been exhaustively researched, it is known that sleep is required for effective rest and repair, consolidation of memory, easing of stress, improved mood, weight control, among many others. It is vital for alertness for the following day, or even the same day. It is recommended that an adult gets 6-8 hours of sleep daily.

However, the quality of sleep each individual gets is affected by a range of factors, both internal (such as sickness) or external (such as food, stress levels, social behaviour like smoking and alcohol drinking).

AIMS & OBJECTIVES.

This project’s major aim is to determine the factors affecting an individual’s sleep efficiency, using machine learning methods and exploratory data analysis. In other words, Team KNN wanted to use this project to describe the trends in sleep efficiency using available data, and to predict future subjects’ sleep efficiency using data modelling and machine learning. Insights could also be applied to real-life subjects.

PROJECT FLOW.

To achieve the set aims and objectives for this project, we followed these listed steps:

* Data Sourcing & Description of Features
* Data Preparation/Cleaning
* Exploratory Data Analysis
* Data Modelling & Evaluation

Data Sourcing & Description of Features

Data Modelling & Evaluation

Data Preparation/Cleaning

Exploratory Data Analysis

DATA SOURCING.

The data was provided to us from a [Kaggle](https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency) source. It contained 452 unique subjects with 15 features such as age, gender, smoking status, exercise frequency, among others. Throughout the project, the python libraries used were Pandas, Numpy, Matplotlib, Seaborn and Scikit-Learn.

Description of Features.

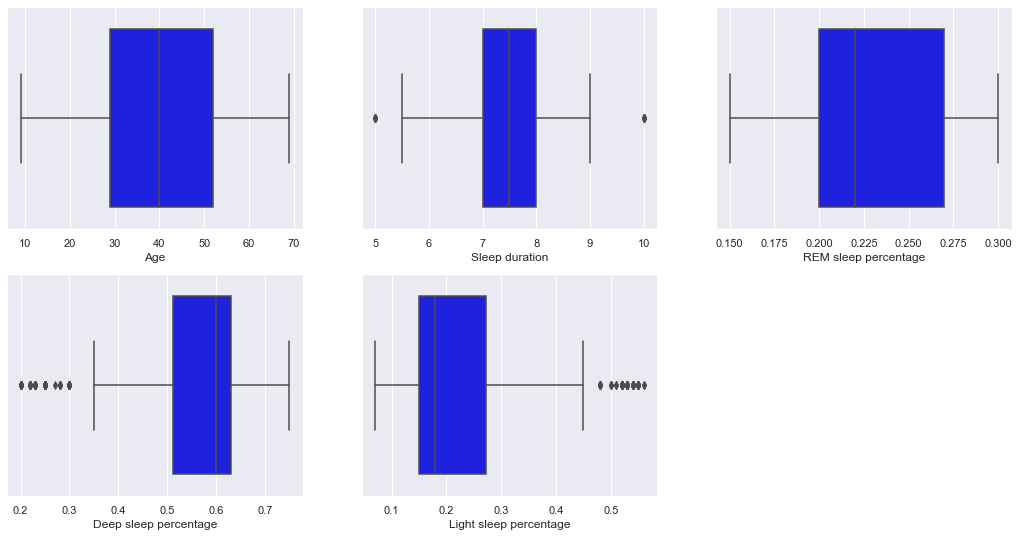
1. ID: Each test subject is identified by a unique identification number
2. Age: Age of each subject.
3. Gender: the gender of each subject (Male or Female).
4. Bedtime and Wakeup time: These indicate when each subject goes to bed and wakes up each day.
5. Sleep duration: This records the total amount of time each subject slept in hours.
6. Sleep efficiency: This is a measure of the proportion of time spent in bed that is actually spent asleep.
7. REM sleep percentage, Deep sleep percentage, and Light sleep percentage features indicate the amount of time each subject spent in each stage of sleep. REM, which stands for Rapid Eye Movement
8. Awakenings: This records the number of times each subject wakes up during the night.
9. Caffeine and Alcohol consumption: These are recorded in the 24 hours prior to bedtime and measured in milligrams(mg).
10. Smoking status: If the subject smokes, it is recorded as YES and NO if he doesn’t.
11. Exercise frequency: The number of times a subject engages in exercise in a week.

*\*All descriptions were obtained from Kaggle.*

DATA PREPARATION/CLEANING.

In this step, we prepared and cleaned the data to make it fit for modelling and exploratory data analysis. The steps taken were:

* Categorical data such as gender and smoking status which were recognized as string/object datatype were converted as required.
* ID, Bedtime and Wakeup time were dropped as they were not required. Bedtime and Wakeup time were already subtracted to calculate the sleep duration.
* Missing data in the Awakenings, Caffeine consumption and Alcohol consumption columns were replaced with 0, as it was considered to be the more appropriate option.
* REM sleep, deep sleep and light sleep percentage which were erroneously represented in integers from 0-100, instead of float values from 0-1, were adjusted.
* For easier manipulation of subjects’ ages, they were distributed into bins: Child, Teen, 20+, 30+, 40+, 50+ and 60+.
* Using box plots, 55 outliers were identified using Z-scores. However, they were allowed to remain as they could be crucial to the creation of a proper model.



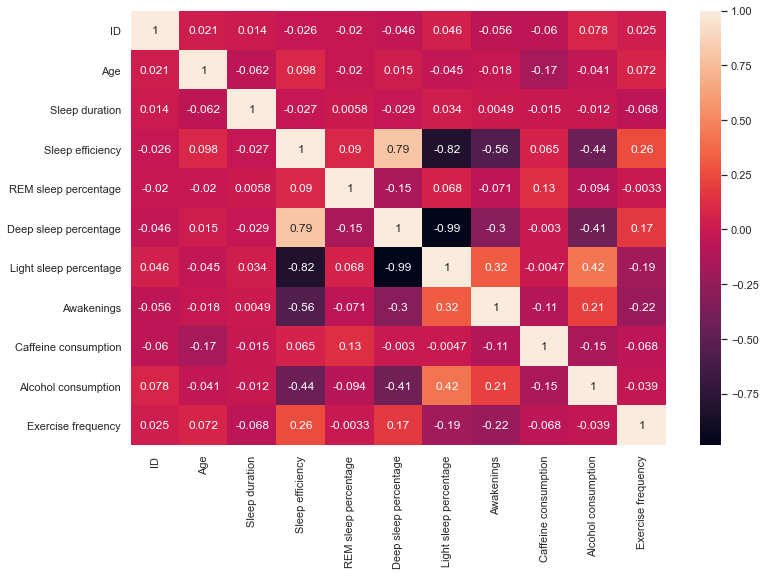
*Box plots showing outliers.*

*\*Subsequently, Sleep Efficiency will be written as SE to avoid redundancy.*

EXPLORATORY DATA ANALYSIS.

On exploratory data analysis, the following insights were generated:

1. Assessment of the distribution of sleep duration showed that most of the subjects slept for 7 hours at least, with 7.46 hours being the average.
2. On average, SE was 78.89%
3. There is no correlation between gender and SE.
4. There is also no correlation between gender and sleep duration as average sleep duration for men was, while it was for women.
5. Furthermore, there is no correlation between gender and other factors like REM sleep, deep sleep, and light sleep percentage.
6. SE in children, lower than that of adults, in spite of sleep duration. This made us conclude that SE may not be quite correlated to sleep duration.
7. It was shown that SE declined with age, peaking at the 30+ age group and declining as subjects in older age groups showed worsening SE.
8. It was also observed that smokers’ average SE was marginally worse than that of non-smokers.
9. Sleep efficiency gradually increased with subjects’ increasing exercise frequency.
10. Fewer awakenings during the night showed improved sleep efficiency.
11. From the correlation plot, it is observed that Deep sleep percentage has a strong positive correlation with our target feature, Sleep efficiency.
12. Also, Light sleep percentage is strongly negatively correlated with Sleep efficiency that Sleep efficiency decreases with increase in Light sleep percentage and increases with an increase as well in Deep sleep percentage.



*Correlation Plot.*

DATA MODELLING & EVALUATION

Data modelling was done using scikit-learn and XGBoost. Using train test split from scikit-learn’s model selection, the test size parameter was set to 0.3, splitting the data into 70% training data and 30% test data. This needed to be done as we were not exactly provided with a separate dataset specifically for training. We selected Sleep Efficiency as our target variable.

Furthermore, we used two training sets: one that used the age bins as categorical data and one that used the raw ages as continuous quantitative data.

The linear modelling methods we attempted using were the Gradient Boosting Regressor, Random Forest Regressor, Linear Regression, ElasticNet, Lasso, Ridge, and Extreme Gradient Boost Regressor. Using statistical values such as the R-squared score, mean absolute error, root mean squared error and residual sum of squares, we compared all these values with the models, and documented them into a dataframe.

Of all the models used, we chose Random Forest Regressor and the scaled dataset of using the ages as continuous data as our best set for the best model scores. This is because it gave 88.9% R-squared score. This tells us that the variance of the data from the mean of Sleep efficiency is reduced by 88.9%.

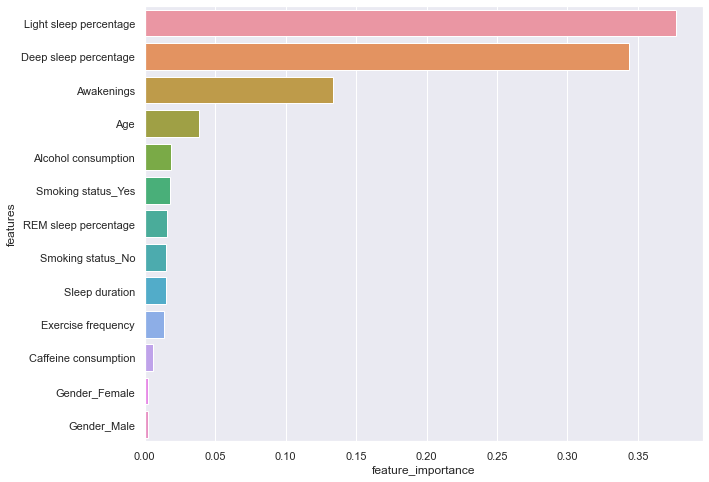
Moreover, with a Root Mean-squared error of 0.0474, it showed that we could peg any subject’s sleep efficiency at ±4.74% if scaled. Comparing with the dataset using the age bins, our RMSE was 0.0482. considering the dataset with the raw age, our RMSE was 0.474 (4.74% from the actual values in the test set).

The model was then saved as a pickle file which can be loaded up and deployed to an endpoint at any required time. It was hosted on [Github](https://github.com/Sammybams/HamoyeAI-Team-KNN-Premier-Project), afterwards.

CONCLUSION.

In summary, using the modelled data and exploratory data analysis, we concluded that the factors responsible for good/poor sleep efficiency are:

1. Awakenings
2. Age
3. Alcohol consumption
4. Smoking status
5. Sleep duration



*Bar Chart Showing the Main Factors Affecting Sleep Efficiency*

Therefore, recommendations could be made. Age is not a factor that can be controlled. However, for the controllable ones:

* One should reduce the number of times they wake up at night. This could be done by urination before going to bed, use of earplugs, ensuring proper silence and darkness and removal of nearby electronic devices.
* Alcohol intake should be gradually reduced or stopped completely if possible.
* Individuals should avoid smoking.
* One should make the effort to sleep more at night. This could be done by increasing one’s exercise frequency and overall activity.

TEAM MEMBERS.

1. Harish Chauhan – Team Lead.
2. Samuel Adurayemi Bamgbola – Assistant Team Lead.
3. Stephanie Netoafomachukwu Anyama – Query Analyst.
4. Eke Mong Eke
5. Olamide James Amoo
6. Aleksander Busz Fabritius
7. Monicah Omondi Atieno
8. Rahul Kumar
9. Omojasola Juwon Ezekiel
10. Jeffrey Ogeh