Class	4
Full Name	Janelle Boey
University Email	JBOEY002@e.ntu.edu.sg
Matriculation Number	U2210772D

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

[X] I have read and accept the above.

Table of Contents

Answer to Q1: Explore the data and show three interesting findings.

Finding 1

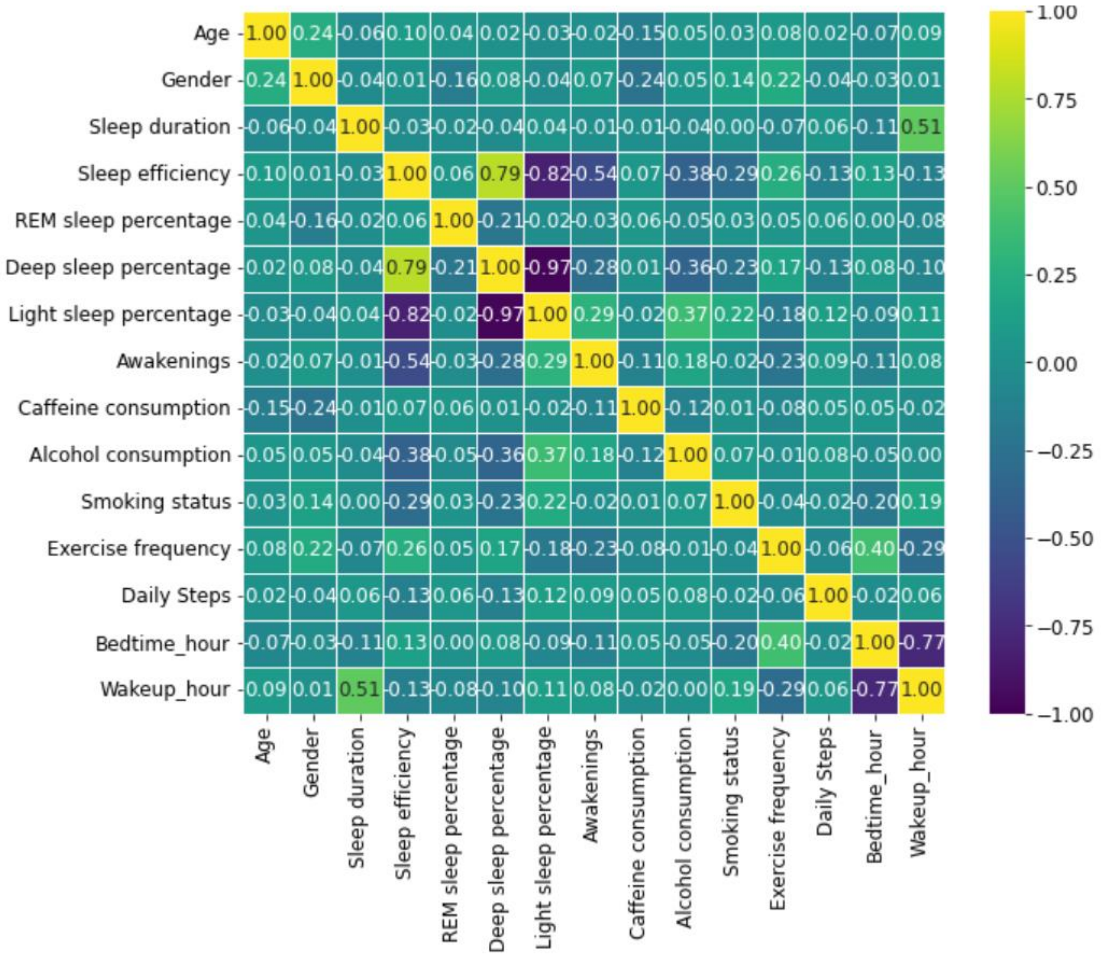


Figure 1: Heatmap of all variables in dataset

Based on the heatmap, there is a significantly high positive correlation between sleep efficiency and deep sleep percentage of 0.79. This means that the higher the deep sleep percentage in the duration of sleep, the higher the sleep efficiency. This shows that deep sleep is the most important stage of sleep to help improve the quality of sleep of an individual. Thus, it can be concluded that factors influence deep sleep percentage would in turn affect sleep efficiency, which will ultimately determine an individual's quality of sleep.

Finding 2

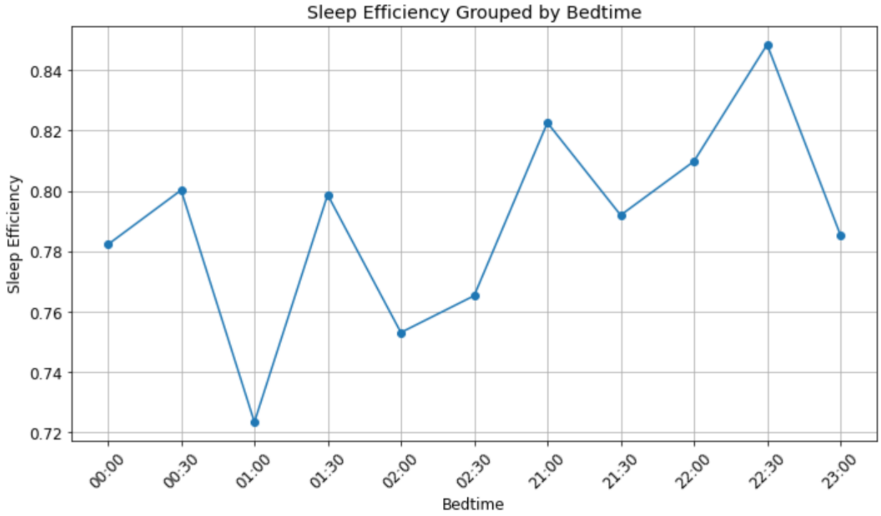


Figure 2: Line graph of sleep efficiency according to bedtime

Generally, sleep efficiency diminishes as individuals sleep later. From the line graph, it can be observed that sleep efficiency ranges from 0.78 to 0.85 for individuals that go to bed by 10:30pm. Whereas, for individuals that go to bed past 10:30pm, the range of sleep efficiency decreases to about 0.72 to 0.80. Notably, the optimal bedtime is 10:30pm, where sleep efficiency is the highest compared to the other bedtimes. Conversely, bedtime at 1am is associated with the worst sleep efficiency.

Finding 3

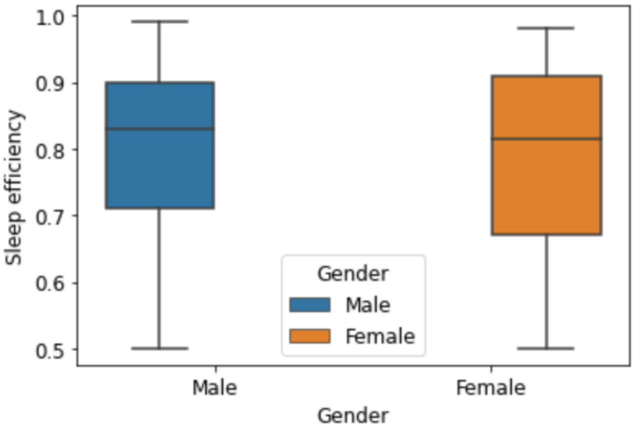


Figure 3: Boxplot of sleep efficiency by gender

Sleep efficiency could be attributed to the demographics of an individual as well. The boxplot depicts that males have a slightly higher median sleep efficiency as compared to females. It can also be noted that females have a wider distribution of sleep efficiency, which indicate that sleep efficiency is more varied for females than males. However, there should be a caveat that the disparity between male and female sleep efficiency could have been attributed to different lifestyle behaviours. For instance, if males tend to exercise more frequently than females, then the males' improved sleep efficiency could have been due to healthier lifestyle habits.

Answer to Q2: Propose two research questions that can be answered from the dataset and models learnt in this course.

Questions:

1. Do patterns of lifestyle habits help to predict sleep quality?
2. Does biological makeup (i.e. Age, Gender) help predict sleep quality?

Answer to Q3: State the target variable and input variables that will be used. Explain your choice.

The target variable “sleep” is a combination of all the sleep metrics which have been normalized into percentages. “sleep” includes “Sleep duration”, ‘REM sleep percentage”, “Light sleep percentage”, “Deep sleep percentage” and “Sleep efficiency”.

The combination of these metrics into a single target variable is to capture a holistic picture of an individual’s sleep quality. Each metric provides information about the different aspects and components of sleep, such as the duration of sleep and proportion of time spent in various sleep stages and how sleep is efficiently achieved. By aggregating these metrics into a single variable “sleep”, it enhances the efficiency of the analysis and interpretation of overall sleep quality.

The input variables would include features that could help provide insights into the research questions i.e. lifestyle and biological makeup. Hence, the input variable “x1” includes, “Age”, “Gender”, “Smoking status”, “Caffeine consumption”, “Alcohol consumption”, “Exercise frequency”, “Bedtime_hour” and “Wakeup_hour”.

The goal of including these input variables is to capture a broad range of lifestyle and biological factors that may influence sleep quality. Analysing their relationships with the target variable “sleep”, can help identify significant predictors and factors that contribute to overall sleep quality. The variable “Daily steps” was omitted to prevent overfitting of the model.

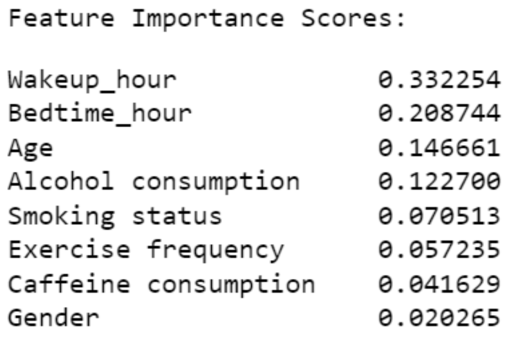
Answer to Q4: Compare the predictive performance of at least 2 types of model learnt in this course for the target variable, and display the results in a table. Which model performed the best? Note: one of the model must be random forest

The models are evaluated based on 3 metrics: R-squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Model	**R-squared**	**RMSE**	**MAE**
Random Forest	0.71	2.27	1.17
Neural Network	0.47	3.08	2.44
Linear regression	0.40	3.29	2.69

Based on the metrics, Random Forest is the model that seems to perform the best. The R-squared value for Random Forest is the highest at 0.71, which suggests that the model is the best fit to measure the proportion of the variance in the dependent variable explained by the independent variables in the model. RMSE for Random Forest is the lowest at 2.27 which suggests that the model's predictions are close to the actual values and there are smaller errors between predicted and actual values. Finally, the MAE for Random Forest is the lowest as well, which indicates that the model performs the best in terms of the average magnitude of errors in the prediction.

Answer to Q5: Answer the two research questions.



```
Feature Importance Scores:

Wakeup_hour              0.332254
Bedtime_hour             0.208744
Age                      0.146661
Alcohol consumption      0.122700
Smoking status           0.070513
Exercise frequency       0.057235
Caffeine consumption     0.041629
Gender                   0.020265
```

Figure 4: Random Forest feature importance scores

Lifestyle Habits

Based on the Random Forest feature importance scores, wake up hour has the highest feature importance score of 0.33. This is then followed by bedtime hour, alcohol consumption, smoking status, exercise frequency and lastly caffeine consumption. The results suggest that wake up hour has the most substantial impact on predictions of overall sleep quality of an individual. Bedtime hour was identified as the 2nd most important variable predicting sleep quality, followed by alcohol consumption. The remaining features which include smoking status, exercise frequency and caffeine consumption seem to have less significance on the predictability of sleep quality.

Hence, lifestyle habits such as wake up hour, bedtime hour and alcohol consumption are the predictors that can be used to predict the quality of sleep. Other features which include, smoking, exercise and caffeine may not be the as useful for the predictive performance of sleep quality.

Demographic

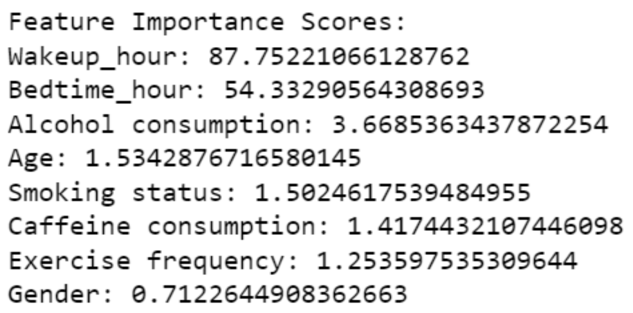
Referencing the feature importance score for age and gender, age seems to have a larger influence on prediction of sleep quality as compared to gender.

Age has the 3rd highest feature importance score among all the features. This indicates that age has one of the most significant influences on the model's predictions. On the other hand, gender is the last and least important feature that influences the model's predictions. Thus, it can be concluded that age can be used to predict the quality of sleep, while gender would most likely not be useful for such predictions.

Answer to Q6: Summarize the most important findings from analytics in less than 500 words.

As seen in the previous question, figure 4 shows that the random forest model predicted that wake up hour has the highest feature importance score and thus most influential in determining predictions for quality of sleep. The importance score then decreases in order of bedtime hour, alcohol consumption, smoking status, exercise frequency and lastly caffeine consumption.

The neural network model suggests a different ranking of feature importance for the same training and test set that were fed into the random forest model.

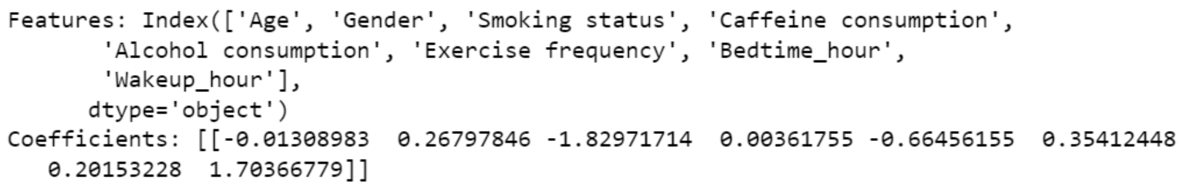


```
Feature Importance Scores:
Wakeup_hour: 87.75221066128762
Bedtime_hour: 54.33290564308693
Alcohol consumption: 3.6685363437872254
Age: 1.5342876716580145
Smoking status: 1.5024617539484955
Caffeine consumption: 1.4174432107446098
Exercise frequency: 1.253597535309644
Gender: 0.7122644908362663
```

Figure 5: Neural Network feature importance scores

The neural network model similarly ranks wakeup and bedtime hour as the top 2 most important features that affect sleep quality predictions. However, as compared to the random forest model, neural network ranks alcohol consumption as the 3rd most important feature before age. Both random forest and neural network ranked smoking status as the 5th most important feature. Neural network determined that caffeine consumption has a higher importance than exercise frequency. Finally, both models ranked gender as the least important feature in predicting sleep quality. Hence, the consistency across both models, show that wakeup hour and bedtime hour are the most important and influential features, while gender has the least impact on predicting sleep quality.

The linear regression model has to be interpreted differently from the feature importance scores in random forest and neural network models. Since linear regression measures correlation between input variables to the target variable, we could measure the extent of each variable in predicting the outcomes of sleep quality based on the weightage of their coefficients.



```
Features: Index(['Age', 'Gender', 'Smoking status', 'Caffeine consumption',
       'Alcohol consumption', 'Exercise frequency', 'Bedtime_hour',
       'Wakeup_hour'],
      dtype='object')
Coefficients: [[-0.01308983  0.26797846 -1.82971714  0.00361755 -0.66456155  0.35412448
   0.20153228  1.70366779]]
```

Figure 6: Linear Regression coefficients

As seen from the weightage of the coefficients produced by the linear regression model, smoking status seems to be the variable that has the strongest correlation to the target variable. This indicates that the frequency of smoking would lead to the most significant increase or decrease in quality of sleep. The variable that has the 2nd strongest correlation to the target variable is wakeup hour, of which both random forest and neural network determined that is the most important feature in predicting sleep quality. This shows that the ranking of feature importance may not necessarily be associated with the correlation between the input variable and the target variable. The weightage of the coefficients then decreases in the order as follows, alcohol consumption, exercise frequency, gender, bedtime hour, age and lastly caffeine consumption.

Answer to Q7: Suggest a way that your model could be used in the real world. Explain

In today's fast-paced society, many individuals experience poorer sleep quality due to various stressors stemming from circumstances and unhealthy lifestyle habits. Despite being one of the most crucial components of overall health and well-being, sleep is often overlooked and compromised on because of other priorities. However, research consistently shows that good sleep quality is vital for improving productivity and enhancing overall quality of life.

To address this growing concern, wearable sleep monitoring gadgets and applications offer a promising solution. These devices allow individuals to conveniently track and monitor their sleep patterns, providing valuable insights into the factors that may be impacting the quality of their sleep.

The model can be integrated into sleep monitoring wearable gadgets such as smart watches, to help improve an individual's quality of sleep. Based on the model, the extent that lifestyle habits and biological factors can influence quality of sleep. Users can thus input information about their lifestyle habits such as alcohol consumption, exercise frequency, smoking status, and caffeine intake. The gadget can track and monitor the bedtime and wake up patterns of the individual and utilize the model to analyse these inputs to predict the user's sleep quality. The gadget will then provide personalized insights on how their lifestyle habits are impacting their sleep and receive personalized feedback to improve their sleep quality, such as suggestions for the user to go to bed earlier, or to maintain a healthy bedtime routine.

This approach empowers users to take proactive steps towards optimizing their sleep habits and ultimately get better sleep quality and improve overall health and wellbeing.