WIE2003 : INTRODUCTION to DATA SCIENCE

SEMESTER 2, 2023/2024

GROUP ASSIGNMENT PART 2

TITLE: SLEEP QUALITY PREDICTION

LECTURER: PROFESOR MADYA DR. MAIZATUL

AKMAR BINTI ISMAIL

CLASS: OCCURENCE 1 GROUP 11

NAME	MATRICS NUMBER
NG ZHI WEI	23051966
LIEW JIN SZE	23005226
KUEH PANG LANG	23005227
SILVIA EVAFARINA ASUN	23062999
ELAINE LEOW HUI XIN	22094003

1. Project background

Our project focuses on analyzing sleep patterns by considering various health and lifestyle factors that impact sleep quality and duration. We are utilizing a comprehensive dataset sourced from Kaggle, containing 374 rows and 13 columns with diverse information such as gender, age, occupation, sleep duration, sleep quality ratings, physical activity levels, stress levels, BMI category, blood pressure, heart rate, daily step counts, and the presence or absence of diagnosed sleep disorders.

This project is highly relevant and suitable for organizations like the National Sleep Foundation, Sleep Research Society, and Consumer Technology Association, which can leverage their expertise, resources, and technologies to achieve common goals of improving sleep health and promoting better lifestyle habits related to sleep. The target audience for this project is quite broad, including researchers and data scientists interested in analyzing sleep patterns, healthcare professionals seeking to better assess and treat sleep disorders, individuals looking to optimize their sleep for improved well-being and productivity, as well as technology developers aiming to innovate new sleep tracking tools and digital health solutions.

A key potential benefit is enhancing overall sleep health for individuals by providing personalized insights derived from analyzing their sleep patterns in relation to various lifestyle factors. This can empower people to make more informed choices towards improving their sleep habits, leading to better overall well-being, increased productivity, and higher quality of life.

Moreover, the findings from this analysis have the potential to drive innovation in consumer sleep technology, particularly in the development of more accurate, user-friendly, and comprehensive sleep tracking devices and digital health solutions focused on sleep. This could revolutionize how individuals monitor and proactively manage their sleep, ushering in a paradigm shift where quality sleep is viewed as a vital component of a healthy lifestyle.

Through our holistic approach that integrates analysis of both physiological sleep metrics and lifestyle factors influencing sleep, this project aims to foster a culture that recognizes and prioritizes sleep as a fundamental pillar of optimal health and performance. By raising awareness, providing education, enabling community engagement, and developing targeted interventions, we strive to bring about positive societal change where quality sleep is embraced and valued as essential for flourishing.

2. Project Objectives

The primary objectives of this project are:

1. To train and evaluate multiple machine learning models on the provided sleep dataset in order to determine the optimal model for accurately predicting sleep quality scores based on the various factors and features present in the data.

2. To conduct comprehensive feature importance analysis and interpretation to identify the most significant factors and variables that influence sleep quality.

3. To develop a data product that leverages the trained predictive model to provide personalized sleep quality predictions and recommendations for users based on their input data and the identified important features impacting sleep quality.

4. To provide personalized sleep insights and guidance through a data product, empowering better sleep quality and well-being.

3. Data Modelling

To predict sleep quality, we explored the performance of four different regression models. Given that sleep quality is a continuous variable, we opted for regression models. Our approach began with the most fundamental regression model, linear regression, and progressed to more sophisticated ensemble methods like decision trees and random forests. To comprehensively evaluate the performance of each regression model, we utilized the following metrics:

1. **R^2 Score (Coefficient of Determination):** This metric quantifies the proportion of the variance in the target variable that is explained by the independent variables in the model. A higher R^2 score indicates a better fit of the model to the data.
2. **Mean Squared Error (MSE):** The MSE measures the average squared difference between the actual and predicted values. Lower MSE values suggest that the model's predictions are closer to the actual values, indicating better predictive accuracy.
3. **Root Mean Squared Error (RMSE):** The RMSE is the square root of the MSE, providing a more interpretable measure of the average prediction error.

3.1 Polynomial Ridge Regression

Polynomial ridge regression is a robust algorithm that combines polynomial feature transformation with ridge regularization to predict sleep quality. It works by fitting a polynomial equation to the data, which allows for capturing non-linear relationships, while ridge regularization helps to prevent overfitting by penalizing large coefficients.

Utilizing the Polynomial features and Ridge modules from the scikit-learn library, we transformed the input features into polynomial features and then applied ridge regression to train the model with the training dataset. This approach helps to manage multicollinearity and improves the model's accuracy and generalization capabilities by balancing the complexity of the polynomial terms with regularization.

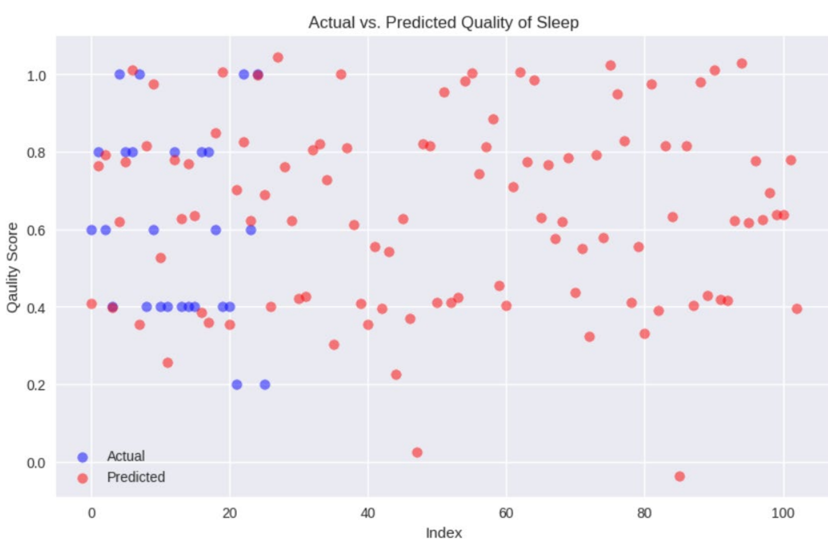
After training the model, we evaluated its performance by using the same dataset for prediction, allowing us to assess its ability to capture the complex relationships within the data accurately.

Upon completing the training and prediction phases, we calculated several evaluation metrics on the training data:

- Mean Squared Error (Training): 0.0023595748616676124
- Root Mean Squared Error (Training): 0.048575455341845354
- R^2 Score (Training): 0.96199330482899365

The performance metrics of the polynomial ridge regression model on the training data indicate a highly accurate and effective predictive capability. The Mean Squared Error (MSE) is exceptionally low at 0.00236, suggesting that the model's predictions are very close to the actual values. This is further supported by the Root Mean Squared Error (RMSE) of 0.0486, indicating that, on average, the predictions deviate from the actual values by only 0.0486 units. Additionally, the R^2 score of 0.962 demonstrates that the model explains 96.2% of the variance in the quality of sleep.

Visualization: Actual vs. Predicted Values



From the plot, we can see that the red points (predicted values) closely follow the blue points (actual values) across most of the index range. This close alignment suggests that our polynomial ridge regression model has a good fit and accurately predicts the quality of sleep for the majority of cases.

However, there are some discrepancies where the predicted values deviate from the actual values, particularly in certain regions of the plot. These deviations highlight areas where the model's predictions are less accurate, possibly due to complexities in the data that are not fully captured by the model or the presence of outliers.

Hyperparameter Tuning

During the hyperparameter tuning phase, we employed the GridSearchCV technique from the scikit-learn library to systematically explore a defined hyperparameter space. After tuning, we obtained the following evaluation metrics:

- **Mean Squared Error: 0.002360**
- **Root Mean Squared Error: 0.048576**
- **R^2 Score: 0.961993**

By fine-tuning the hyperparameters, we aimed to enhance the model's predictive accuracy and generalization capability. The results of this tuning process will be instrumental in determining the effectiveness of the polynomial ridge regression model in predicting sleep quality.

3.2 Support Vector Regression

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. The SVR algorithm aims to find the hyperplane that passes through as many data points as possible within a certain distance, called the margin.

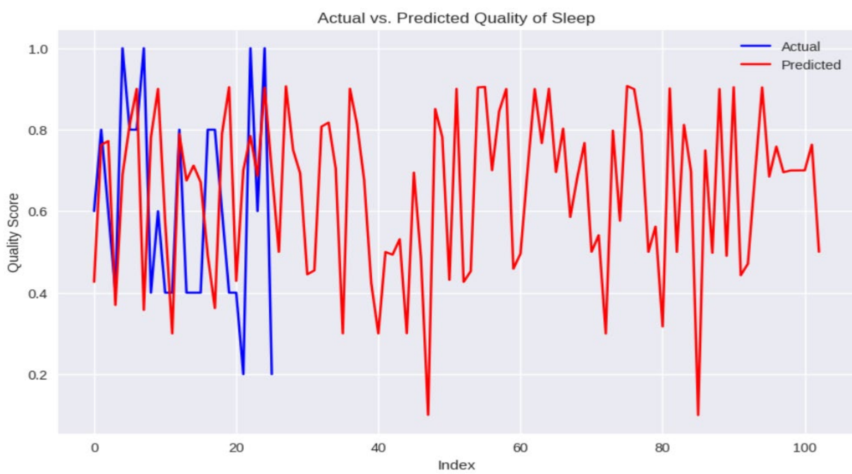
First, we utilize the SVR module from scikit-learn library for model training with kernel='rbf' which specifies the radial basis function (RBF) kernel and is commonly used in SVR because of its flexibility in capturing complex relationships in the data, C=1.0 as the regularization parameter to control the trade-off between maximizing the margin and minimizing the training error, where a larger value of C implies a smaller margin but less error penalty, potentially leads to overfitting, epsilon=0.1 specifies the margin of tolerance where no penalty is given to errors within this distance. It allows the model to fit the data within a certain error margin.

Upon completing the training and prediction phases, we calculated several evaluation metrics on the training data which are:

- **Mean Squared Error (Training): 0.005954647514724557**
- **Root Mean Squared Error (Training): 0.07716636258580908**
- **R^2 Score (Training): 0.9040859111446119**

The provided evaluation metrics for the training data indicate that the Support Vector Regression (SVR) model performs well. With a mean squared error (MSE) of approximately 0.0059 and a root mean squared error (RMSE) of around 0.077, the model's predictions are close to the actual values. Additionally, the R^2 score of approximately 0.904 suggests that the model explains about 90.4% of the variance in the training data, indicating a good fit. Overall, these metrics indicate that the SVR model effectively captures the relationship between the features and the target variable in the training set.

Visualization: Actual vs Predicted Values



By using a Support Vector Regression (SVR) model to predict sleep quality, the comparison of actual recorded scores (blue line) to the SVR predictions (red line), shows that the predictions are erratic and often miss the mark. However, as we continue, the SVR model begins to follow the general trend of sleep patterns, even if it's not always precise. In the later stages, while the predictions remain variable, they generally align with the overall pattern of sleep quality. This indicates that the SVR model has potential but still requires refinement for better accuracy.

Hyperparameter Tuning

Hyperparameters can be tuned to improve the performance of an SVR model. In the hyperparameter tuning process for SVR, we define a parameter grid containing the settings for key hyperparameters which are kernel type, regularization strength and epsilon value. Utilizing grid search with cross-validation, we systematically explore these combinations to find the optimal set that minimizes the mean squared error on our training data. Finally, we select the combination that yields the best performance and print out its hyperparameters, enabling us to fine-tune our SVR model for improved predictive accuracy.

After tuning, we obtained the following evaluation metrics:

- **Mean Squared Error: 0.002796**
- **Root Mean Squared Error: 0.052880**
- **R^2 Score: 0.954959**

3.3 Decision Tree Regression

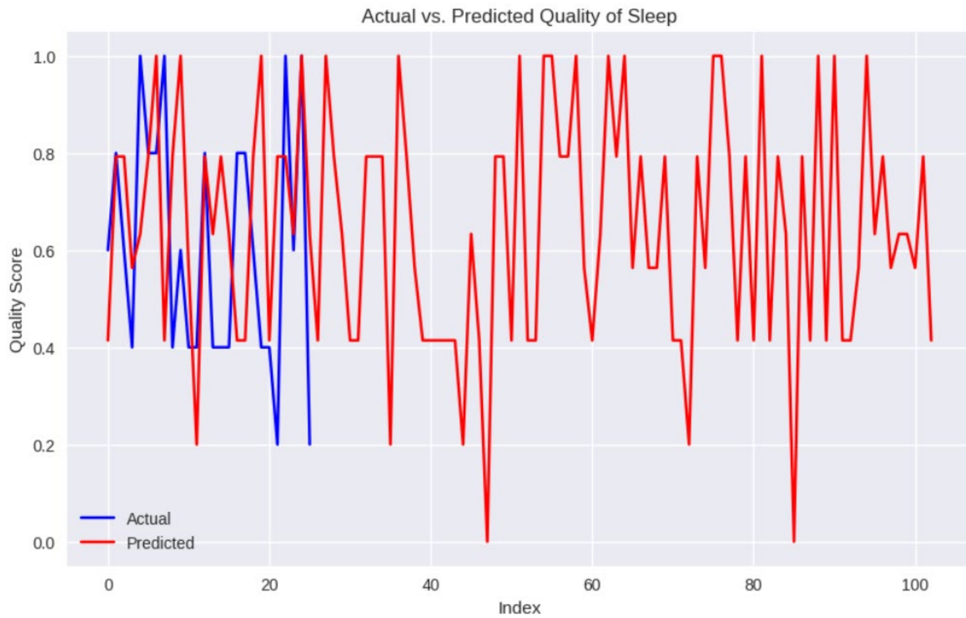
First, we utilize scikit-learn's Decision Tree Regressor to build a decision tree regression model with a maximum depth of 3. This is to provide a balance between model complexity and overfitting. By fitting the model to the training data and making predictions, we evaluate its performance using the key metrics: MSE, RMSE and R^2. These metrics gauge the model's ability to accurately predict the target variable and its goodness of fit to the training data. Adjusting the max depth parameter allows us to control the complexity of the decision tree, influencing its ability to generalize to unseen data.

Upon completing the training and prediction phases, we calculated several evaluation metrics on the training data which are:

- Mean Squared Error (Training): 0.0031551296599840284
- Root Mean Squared Error (Training): 0.056170540855363216
- R^2 Score (Training): 0.9491789588200374

From this result, we can see that the obtained Mean Squared Error (MSE) of 0.00315 and Root Mean Squared Error (RMSE) of 0.0562 signify minimal prediction errors, while the R-squared (R^2) score of 0.949 indicates that the model explains approximately 94.9% of the variance in the target variable, suggesting a robust fit to the training data.

Visualization: Actual vs Predicted Values



Using a Decision Tree Regression model to predict sleep quality, the early predictions are inconsistent, often missing the mark. As you continue tracking, the model begins to align with the general trend of sleep patterns, but not always precise. In later stages, predictions remain variable but generally follow actual sleep quality patterns. This indicates that while the model captures broader trends, it needs refinement for more accurate predictions. This highlights the model's potential and areas for improvement.

Hyperparameter Tuning

We utilize scikit-learn's Grid Search CV to systematically search for the optimal hyperparameters of a Decision Tree Regressor model. It defines a parameter grid containing hyperparameters to tune, such as maximum depth and criterion. After finding the best hyperparameters, a new regressor is trained with these parameters.

After tuning, we obtained the following evaluation metrics:

- **Mean Squared Error: 0.000777**
- **Root Mean Squared Error: 0.027869**
- **R^2 Score: 0.987489**

3.4 Random Forest Regression

Recognizing the limitations of a single decision tree, we further enhanced our predictive model by employing random forest regression. It works by creating a multitude of decision trees on different subsets of the data and then averaging the predictions of these individual trees to obtain the final prediction.

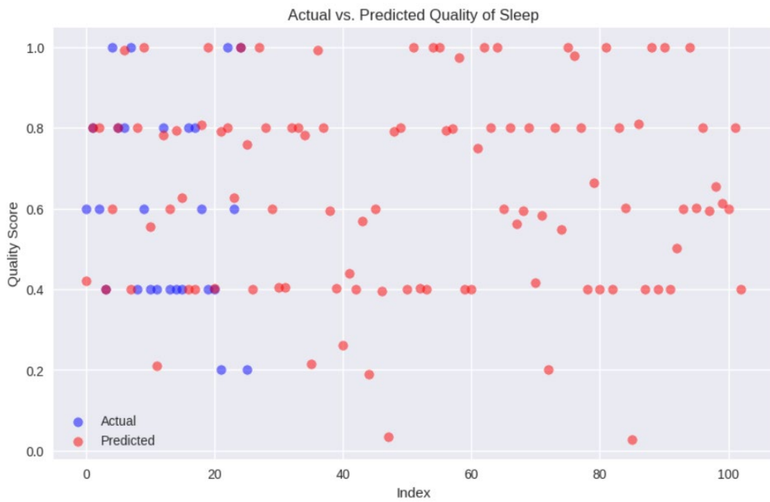
Utilizing the Random Forest Regressor module from the scikit-learn library, we trained the model with the training dataset. Subsequently, we utilized the same dataset for prediction to evaluate the model's performance.

Upon completing the training and prediction phases, we calculated several evaluation metrics on the training data:

- Mean Squared Error (Training): 0.0003972038834951462
- Root Mean Squared Error (Training): 0.019929974498105765
- R^2 Score (Training): 0.9936020648609255

The mean squared error (MSE) on the training dataset is very small. This indicates that, on average, the model's predictions are quite close to the actual values. The Root Mean Squared Error (RMSE) is approximately 0.0199 units. This suggests that the model's predictions are generally within a very narrow range of the actual values. The R^2 Score (Coefficient of Determination) of this model explains approximately 99.36% of the variance in sleep quality using the features included in the model. This high R^2 score indicates that the model captures almost all of the variation in sleep quality.

Visualization: Actual vs. Predicted Values



The scatter plots allow us to observe the relationship between the actual sleep quality values and the corresponding predicted values generated by our model. Each data point represents an observation, with the x-coordinate denoting the actual sleep quality value and the y-coordinate representing the predicted value.

Upon examining the scatter plots, we observe that most predicted sleep quality scores closely align with the actual values. This alignment suggests that our model's predictions are highly accurate, with deviations from the actual values being minimal.

Hyperparameter Tuning

In this phase of hyperparameter tuning, we utilized the Grid Search CV technique from the scikit-learn library to systematically search through a specified hyperparameter space and we obtained the following evaluation metrics for the tuned Random Forest Regression model:

- **Mean Squared Error (MSE):** 0.000373
- **Root Mean Squared Error (RMSE):** 0.019308
- **R^2 Score:** 0.993995

An MSE value of 0.00037 is very small, indicating that the predicted values are very close to the actual values on average. An RMSE of 0.01931 is also very low, suggesting that the model's predictions are quite accurate. An R^2 score of 0.99399 means that the Random Forest model explains around 99.4% of the variance in the target variable.

Overall, the hyperparameter tuning process using Grid Search CV has significantly improved the performance of the Random Forest Regression model in predicting sleep quality.

3.5 Determine the best modal

Lists of MSE, RMSE and R2 scores for different models

	MAE	Train RMSE	R2-Score
Polynomial Ridge Regression	0.002360	0.048576	0.961993
Support Vector Regression	0.002796	0.052880	0.954959
Decision Tree Regression	0.000777	0.027869	0.987489
Random Forest Regression	0.000373	0.019308	0.993995

Among the models evaluated, Random Forest shows the best performance based on the provided metrics. Random Forest has the lowest values for MSE and RMSE indicating that it has the smallest average difference between the predicted sleep quality values and the actual values. This implies that Random Forest achieves better accuracy in predicting sleep quality compared to the other models.

Additionally, Random Forest has the highest R^2 score, which measures the proportion of variance in the target variable (sleep efficiency) that can be explained by the model. A higher R^2 score indicates a better fit to the data. In this case, Random Forest achieves an R^2 Score of 0.99399, suggesting that it explains approximately 99.4% of the variance in sleep efficiency.

3.6 Model Testing

Since the Random Forest Regression model emerged as the best-performing model during our evaluation process, we utilized it to assess the performance on the testing set for predicting sleep quality. The tuned Random Forest Regression model demonstrates exceptional performance on the testing set which show in the following:

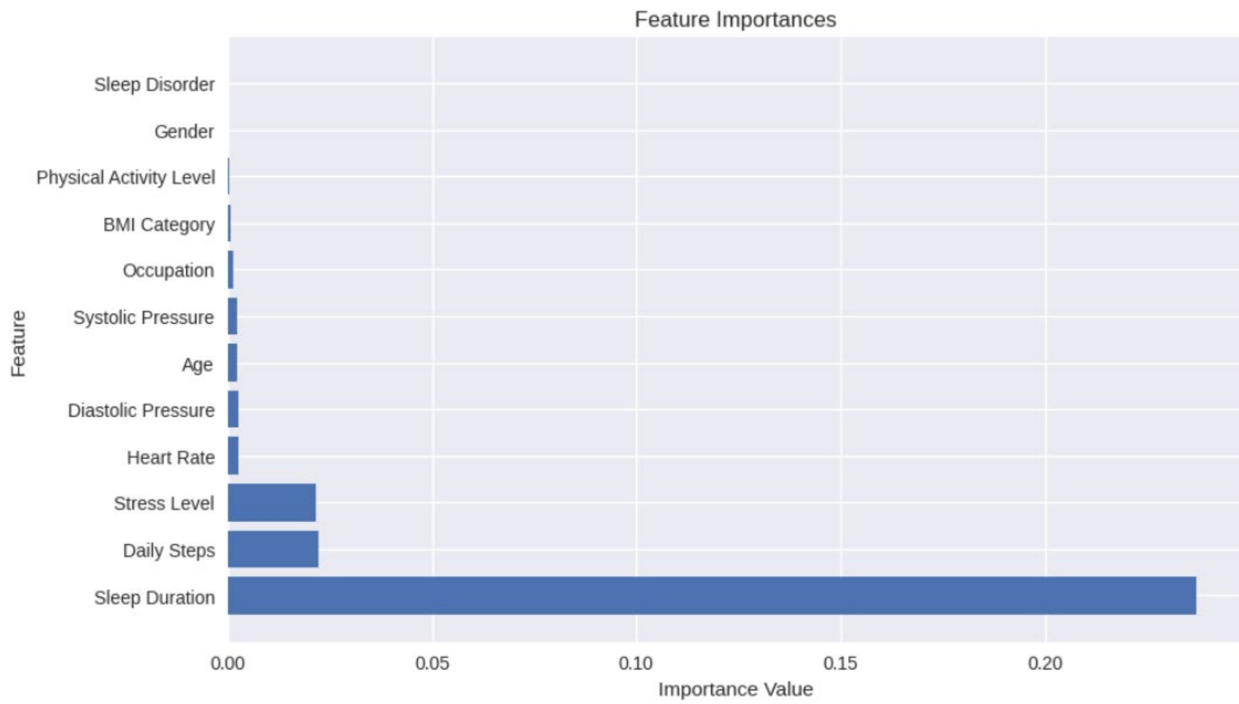
Tuned Random Forest Regression - Testing Set

- **Mean Squared Error: 0.00470**
- **Root Mean Squared Error: 0.06859**
- **R^2 Score: 0.92152**

On the testing set, the tuned Random Forest Regression model performs exceptionally well in predicting the quality of sleep. The model predicts sleep efficiency values with minimal error, as evidenced by its low mean squared error of 0.00470. The root mean squared error of 0.06859 supports this. Additionally, the R^2 score of 0.921520 indicates that the model has a high degree of predictive power, accounting for 92.15% of the variability in the sleep efficiency data. Overall, the tuned Random Forest Regression model performs well in predicting sleep efficiency on the testing set. Its high R^2 score and low error metrics suggest that it can predict sleep efficiency values with accuracy and reliability. The combination of a high R2-score and low MAE on both datasets reinforces the model's robustness and generalization ability, making it a reliable choice for this prediction task.

	MAE	R2-Score
Train	0.000373	0.993995
Test	0.004700	0.921520

3.8 Data Interpretation

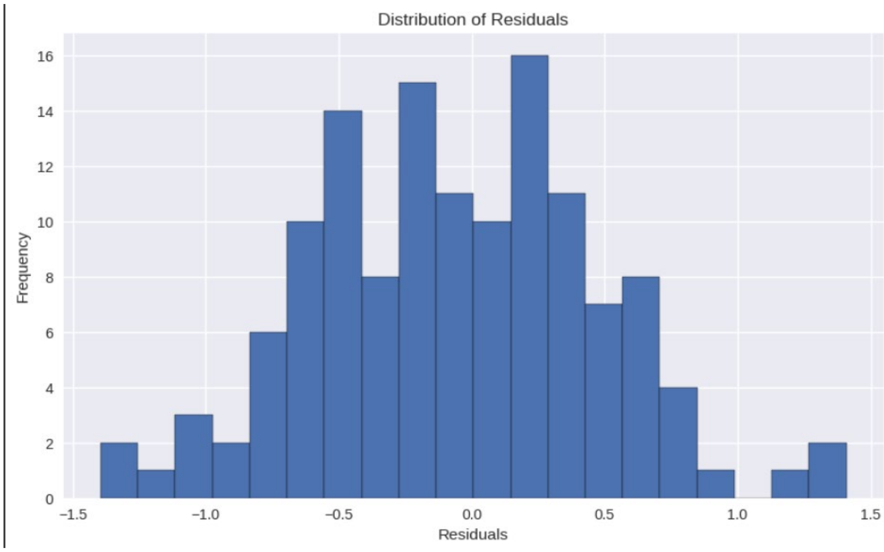


Feature importance

Before modeling, factors such as stress level, BMI category, and possible sleep problems were believed to have a significant impact on sleep quality prediction, but feature importance analysis shows that BMI category and sleep disorder column do not play an important role in determining the sleep quality. The analysis of feature importance shows that sleep duration is the most significant predictor, with a value of 0.236874, which outperforms other variables in terms of its influence on targeted sleep quality. This highlights the vital role of adequate sleep duration in overall health and well-being. Following sleep duration, daily steps emerge as an second important factor, with a feature importance value of 0.022277, highlighting the benefits of regular physical activity in maintaining good sleep quality. Stress level, with a feature importance value of 0.021600, closely follows daily steps, indicating higher stress levels are likely associated with poorer sleep quality.

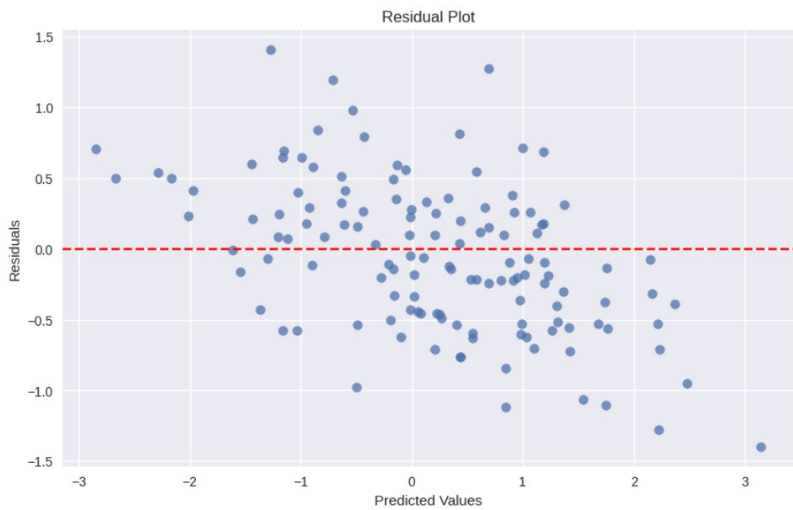
While stress level aligns with prior expectations, the relatively low importance of BMI category and the absence of sleep disorder feature indicate that other factors, such as sleep duration, daily steps, and overall stress management, may be more critical in predicting and improving sleep quality based on this dataset and modeling approach. One possible reason for this could be that BMI category and the presence of sleep disorders are indirectly represented through other variables like stress level and sleep duration. For instance, individuals with higher stress levels or shorter sleep durations might also have higher BMI or sleep disorders, thereby making these indirect predictors less prominent in the analysis.

Residual Analysis



The distribution appears to be approximately bell-shaped, which is a characteristic of a normal distribution. However, there are slightly more negative residuals than positive residuals, suggesting that the model may be slightly overestimating the observed values on average. Additionally, there are a few outliers or extreme values on both the positive and negative sides of the distribution.

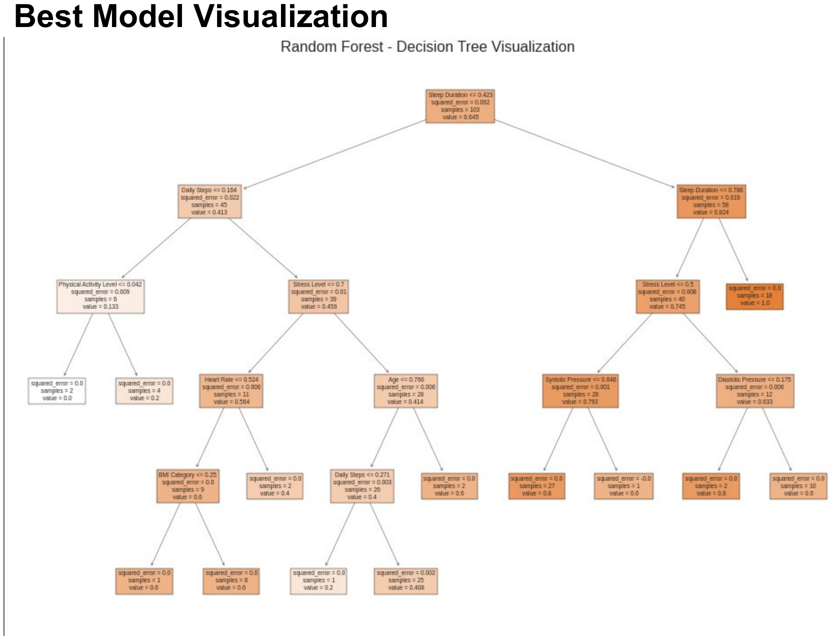
Residual Plot



The residuals appear to be randomly scattered around the zero line, without any clear pattern or trend in relation to the predicted values. This suggests that the assumption of linearity is reasonably met, as there is no obvious non-linear relationship between the residuals and the predicted values. There are a few potential outliers or large residuals, both positive and negative, which are points that deviate significantly from most of the residuals. These outliers could be influential data points or potential issues that may require further investigation.

The residuals are not perfectly symmetrical around the zero line, as there appears to be a slightly higher concentration of positive residuals (overestimates) compared to negative residuals (underestimates). This could indicate a potential bias in the model, where the predictions tend to be slightly higher than the actual observed values on average.

Best Model Visualization



Link for our data modelling and interpretation parts:
https://github.com/zhiweing/WID2003-INTRO-TO-DS-/blob/main/%E2%80%9C_Sleep_Health_Analysis_ipynb%E2%80%9D.ipynb

4.0 Our data product

Our data product titled "Sleep Health Analysis" is a web-based interactive tool built using Shiny, a framework for building interactive web applications in R. The app aims to provide users with insights into their sleep health based on several key health metrics and lifestyle factors. By inputting the user's own data, they can receive personalized feedback on their sleep quality and get suggestions for improvement.

4.1 Our chosen features for user input

The features selected for user input are based on their significant impact on sleep quality. Out of nine potential features, six were chosen due to their strong correlation with sleep health as listed below

1. **Sleep Duration**:
    - **Importance**: Sleep duration is one of the most critical factors affecting sleep quality. Both insufficient and excessive sleep can lead to poor sleep quality and associated health issues.
    - **User Input Method**: Slider to specify the average hours of sleep.
    - **Impact**:
        - Less than 6 hours: May lead to poor cognitive function, mood disturbances, and health problems.
        - 6-8 hours: Generally considered optimal for most adults, leading to good sleep quality.
        - More than 8 hours: Could indicate excessive sleep, potentially linked to underlying health issues.

2. **Daily Steps**:

- **Importance**: Physical activity levels, often measured by daily steps, are closely linked to sleep quality. Regular exercise can improve sleep duration and quality.
- **User Input Method**: Slider to estimate daily steps.
- **Impact**:
    - Less than 5000 steps: May be associated with a sedentary lifestyle, often leading to poor sleep quality.
    - 5000-10000 steps: Indicates moderate physical activity, which can enhance sleep quality.
    - More than 10000 steps: Reflects an active lifestyle, generally beneficial for sleep, though overexertion can occasionally negatively impact sleep.

3. **Stress Level**:
    - **Importance**: Stress levels are directly related to sleep quality. High stress can lead to difficulty falling asleep and staying asleep.
    - **User Input Method**: Slider to rate stress level from 0 (low) to 5 (high).
    - **Impact**:
        - Low stress (0-2): Generally associated with better sleep quality.
        - Moderate stress (3-4): May begin to affect sleep quality.
        - High stress (5): Often linked to poor sleep quality and disturbances.

4. **Age**:
    - **Importance**: Age affects sleep patterns and requirements. Different age groups have varying sleep needs and common sleep-related issues.
    - **User Input Method**: Numeric input for age.
    - **Impact**:
        - Below 18 years: Younger individuals often need more sleep for optimal health and development.
        - 18-60 years: Adults typically require 7-9 hours of sleep for good health.
        - Above 60 years: Older adults may experience changes in sleep patterns and often face sleep disturbances.

5. **BMI (Body Mass Index)**:

- **Importance**: BMI is an indicator of body weight relative to height. It can influence sleep quality, with both high and low BMI associated with different sleep issues.
- **User Input Method**: Numeric input for BMI.
- **Impact**:
  - Underweight (BMI < 18.5): Can be linked to health problems that affect sleep.
  - Normal weight (BMI 18.5-24.9): Generally associated with better sleep quality.
  - Overweight/Obese (BMI ≥ 25): Increased risk of sleep disorders such as sleep apnea.

6. **Heart Rate**:
   - **Importance**: Resting heart rate is an indicator of overall cardiovascular health. It can influence sleep quality, with abnormal heart rates often linked to sleep disturbances.
   - **User Input Method**: Numeric input for heart rate in beats per minute (bpm).
   - **Impact**:
     - Low heart rate (< 60 bpm): May indicate good cardiovascular health, generally beneficial for sleep.
     - Normal heart rate (60-100 bpm): Typically associated with good sleep quality.
     - High heart rate (> 100 bpm): May indicate stress or other health issues affecting sleep.

The features chosen were selected for their strong correlation with sleep quality, as each has a well-documented impact on sleep health. These features cover a broad range of physiological, lifestyle, and psychological aspects that influence sleep. They are easy for users to input through intuitive controls like sliders and numeric inputs, making our web app more user-friendly. Moreover, they provide clear, actionable insights for users to improve their sleep quality, such as increasing daily steps or managing stress levels.

4.2 Our output

The output of the "Sleep Health Analysis" program provides users with a comprehensive analysis of their sleep quality based on the input data they provide. Here's a detailed explanation of each component of the output:

1.

2. **Sleep Quality Text**:

   This section presents a detailed analysis of each input factor (sleep duration, daily steps, stress level, age, BMI, and heart rate) and its relationship with sleep quality. Users receive personalized feedback on how each factor contributes to their overall sleep health. For example, it may highlight the impact of insufficient sleep duration, high stress levels, or abnormal BMI on sleep quality.

3. **Overall Sleep Quality**:

   This part provides users with an overall assessment of their sleep quality based on the input factors. The assessment is typically categorized into three levels: poor, fair, or excellent. The determination is made based on the number of input factors that fall outside the normal or optimal range for promoting good sleep quality. For instance, if most input factors indicate suboptimal conditions (e.g., short sleep duration, high stress, and abnormal BMI), the overall sleep quality may be rated as poor.

4. **Suggestions for Improvement**:

   In this section, users receive actionable suggestions tailored to their specific input data. The suggestions offer strategies and tips to help users improve their sleep quality based on identified areas of concern. For example, if the analysis indicates insufficient sleep duration, the suggestions may include recommendations to establish a consistent bedtime routine, avoid caffeine before bedtime, or create a relaxing sleep environment. Similarly, if stress levels are high, the suggestions may include stress management techniques such as mindfulness meditation, deep breathing exercises, or stress-reducing activities.

Overall, the output of the program aims to empower users with valuable insights into their sleep habits and quality. By providing a detailed analysis of key factors affecting sleep and actionable recommendations for improvement, the program enables users to make informed decisions and take steps toward achieving better sleep health and overall well-being.

Our link for data product:

https://pang-lang.shinyapps.io/SleepHealthAnalysis/

5. Insight and Conclusion

The feature importance analysis revealed surprising insights into the factors influencing sleep quality. Contrary to our initial expectations, sleep duration emerged as the most critical predictor, followed by daily steps and stress levels. These findings challenge our preconceived notions and underscore the importance of a data-driven approach in uncovering meaningful patterns within the data.

Our analysis demonstrated that while variables such as BMI and indicators of sleep disorders did not rank among the top predictors, their influence should not be discounted. It is essential to recognize that these factors could still play a significant role in sleep quality through intricate relationships and potential interactions with other variables. Further research is necessary to delve deeper into these complexities, which could reveal more nuanced insights into how different factors collectively impact sleep quality.

To enhance our understanding and improve predictive models, it is crucial to incorporate additional relevant features such as sleep habits, environmental factors, and physiological measurements. These additional variables could provide a more comprehensive view of the determinants of sleep quality and lead to more robust predictive models. For instance, tracking sleep habits like bedtime consistency, caffeine intake, and screen time before bed can offer insights into behavioral patterns that influence sleep quality. Environmental factors such as room temperature, light exposure, and noise levels can significantly impact sleep.Physiological measurements, such as heart rate variability, body temperature, and respiratory rate, can provide a deeper understanding of the body's internal state and its relationship to sleep quality.

Importantly, our analysis identified the Random Forest model with hyperparameter tuning as the best-performing approach for predicting sleep quality. This model achieved an impressive

R2 score of 0.994 and a low mean absolute error (MAE) of 0.000373 on the training set, demonstrating its ability to capture the complex relationships within the data accurately. Even on the test set, the model maintained a respectable R2 score of 0.919 and a MAE of 0.004838, reinforcing its generalization capabilities. The high performance of the Random Forest model underscores its robustness and reliability in predicting sleep quality.

In conclusion, the feature importance analysis and the application of advanced machine learning models have provided valuable insights into the factors influencing sleep quality. By prioritizing a data-driven approach and continuously refining our models with additional relevant features, we can enhance our understanding of sleep quality determinants and develop more accurate predictive tools. This ongoing research is vital for uncovering the complex interplay of factors affecting sleep quality and ultimately improving interventions and recommendations for better sleep health.