CAN THO UNIVERSITY SOFTWARE CENTER

MEKONG DELTA - APTECH

PROJECT – ANALYZING DATA WITH R

Classifying Sleep Health and Lifestyle

Instructor:

Lam Chi Nguyen

Class:

Students:
1. L22013 – Vo Nguyen Khoa
2. L22016 – Nguyen Thi Kim Cuong
3. L22017 – Tran Thi Cam Lai

Can Tho, 09/2023

TABLE OF CONTENTS

CHAPTER 1: PROBLEM DEFINITION OF PROJECT 3

1.1 Problem 3

1.2 Purpose 4

CHAPTER 2: REQUIREMENTS SPECIFICATION 5

2.1 PROBLEM DEFINITION 5

2.2 SPECIFIC REQUIREMENTS 5

CHAPTER 3: PROJECT PLAN 6

3.1 PROJECT DETAILS 6

3.1.1. Date of the project plan 6

3.1.2. Project vision/objective 6

3.1.3. Scope 6

3.2 KEY STEPS 6

3.2.1. QUALITY PLAN 7

CHAPTER 4: TASK SHEET 9

CHAPTER 5: EXPLORATORY DATA ANALYSIS 11

5.1 DATA SOURCES 11

5.2 DATA DESCRIPTION 11

5.2.1. Gender 11

5.2.2. Age 12

5.2.3. Occupation 12

5.2.4. Sleep Duration 13

5.2.5. Quality of sleep 13

5.2.6. Physical Activity Level 14

5.2.7. Stress Level 14

5.2.8. BMI Category 15

5.2.9. Blood Pressure 16

5.2.10. Heart Rate 17

5.2.11. Daily Steps 18

5.2.12. Sleep Disorder 18

5.3 VISUALIZATION AND BASICS EVALUATION (DESCRIPTIVE ANALYTICS) 20

5.3.1. The relationship between gender and age affect insomnia and sleep apnea? 20

5.3.2. The relationship between occupation and sleep duration affect sleep quality? 20

5.3.3. The relationship between Stress levels affect Heart Rate 22

5.3.4. The relationship between physical activity and daily steps impact Quality of Sleep and Sleep Disorder 24

5.3.5. The relationship between BMI categories affecting sleep quality 25

5.3.6. The relationship between blood pressure and heart rate affect insomnia and sleep apnea 27

CHAPTER 6: AI/ML MODEL SPECIFICATION 30

6.1 MODEL NAME 30

6.2 MODEL SPECIFICATION 30

6.3 DATASET USED FOR MODEL 30

6.3.1. HANDLE MISSING VALUE 30

6.3.2. DATASETS USED FOR MODELING 30

6.4 PROCESS AND CODE ANALYSIS 32

6.4.1. Decision Tree 32

6.4.2. Random Forest 38

6.5 Validation 43

6.5.1. Test validation of the Decision Tree 43

6.5.2. Test validation of the Random Forest 44

CHAPTER 7: CONCLUSION 46

FIGURES

Figure 1. Gender Ratio 11
Figure 2. Age Distribution..12
Figure 3. Count Occupation 12
Figure 4. . Sleep Duration Distribution 13
Figure 5. Quality of Sleep Distribution 13
Figure 6. Physical Activity Level Distribution 14
Figure 7. Stress Level Distribution..14
Figure 8. Count BMI 15
Figure 9. Systolic Distribution 16
Figure 10. Diasotlic Distribution..16
Figure 11. Heart Rate Distribution 17
Figure 12. Daily Steps Distribution..18
Figure 13. Sleep Disorder Ratio 18
Figure 14. Relationship Between Gender and Age 20
Figure 15. Distribution of Stress Level across Sleep Disorder 22
Figure 16. Distribution of Heart Rate across Sleep Disorder....................................22
Figure 17. Relationship between Stress Level, Heart Rate and Sleep Disorder 23
Figure 18. The relationship between Physical Activity, Daily Steps impact Quality of Sleep and Sleep Disorder..24
Figure 19. . Count BMI Categories 25
Figure 20.Relationship between BMI Category, Quality of Sleep and Sleep Disorder.....26
Figure 21. Relationship between Heart Rate, Systolic(Blood Pressure) and Sleep Disorder 27
Figure 22. Relationship between Heart Rate, Díatolic(Blood Pressure) and Sleep Disorder 28
Figure 23. Decision tree 34
Figure 24. X-Val Relative Error..36
Figure 25. Decision tree after prune 37
Figure 26. Random forerst Error 40
Figure 27. Tune Random Forest Parameters 42

TABLES

Table 1. Key steps 7
Table 2. Task sheet 10
Table 3. Datasets used for modeling 32

CHAPTER 1: PROBLEM DEFINITION OF PROJECT

1.1 Problem

Today, life is increasingly developing, along with the appearance of many human diseases. Due to work pressure and life pressure, many of us have encountered sleep disorders.

Sleep disorders can have a significant impact on a person's life, affecting physical and mental health and overall quality of life. Here are some common sleep disorders and their effects:

1. Insomnia: Insomnia is characterized by difficulty falling asleep, staying asleep, or experiencing non-restorative sleep. It can lead to daytime fatigue, decreased concentration, irritability, and impaired cognitive function. Chronic insomnia can also increase the risk of developing mental health disorders such as depression and anxiety.
2. Sleep apnea: Sleep apnea is a condition in which breathing repeatedly stops and starts during sleep. It can cause loud snoring, disrupted sleep, and excessive daytime sleepiness. Sleep apnea is linked to an increased risk of high blood pressure, heart disease, stroke and diabetes. It can also lead to poor concentration, memory problems and mood disorders.
3. Restless Legs Syndrome (RLS): RLS is characterized by an irresistible urge to move the legs, often accompanied by discomfort. It often gets worse during periods of rest or inactivity, making it difficult to fall asleep or stay asleep. RLS can lead to chronic sleep deprivation, daytime fatigue, and impaired quality of life.
4. Narcolepsy: Narcolepsy is a neurological disorder that affects the brain's ability to regulate sleep-wake cycles. People with narcolepsy experience excessive daytime sleepiness, sudden and uncontrollable bouts of sleepiness, and cataplexy (sudden loss of muscle tone). These symptoms can disrupt daily activities, reduce the ability to concentrate, and increase the risk of accidents.
5. Rhythm Disorders : Circadian rhythm disorders occur when a person's internal clock is out of sync with the external environment. Conditions such as shift work disorder and jet lag can disrupt the natural sleep-wake cycle, leading to sleep disturbances, fatigue, decreased alertness and difficulty concentrating.

Sleep disorders can also affect relationships, work performance, and overall quality of life. They can contribute to mood disorders, such as depression and anxiety, and increase the risk of accidents and injuries.

1.2 Purpose

Based on the hospital's examination and treatment results..., including health indicators, stress levels, sleep status and sleep disorders of the patient.

The goal of the study is to understand and evaluate the impact of health indicators and stress levels on sleep disorders. This research could focus on determining the relationship between factors such as physical activity, psychological status, stress levels and health indicators such as blood pressure, sleep quality and symptoms. present sleep disorder. The results of this study may provide important information to better understand sleep disorders and suggest effective treatment and management methods.

CHAPTER 2: REQUIREMENTS SPECIFICATION

2.1 PROBLEM DEFINITION

Sleep disorder is a condition in which a person has difficulty getting or maintaining enough and quality sleep. It can negatively affect health such as making the patient feel tired, lethargic, have no energy, pale skin, loss of appetite and loss of appetite. Sleep disorders also greatly affect the patient's mental state, such as they are often irritable, anxious, sad, or they are lazy in reacting to everyday situations. This can lead to a high chance of the patient suffering from depression, cardiovascular disease or stroke. With different genders and different occupations, sleep disorders are increasing in the current world context.

This study's main purpose is to study the influence of indicators such as gender, age, occupation, sleep quality, sleep duration, physical activity level, stress level, BMI index. , blood pressure, heart rate, number of daily steps to insomnia and sleep apnea.

2.2 SPECIFIC REQUIREMENTS

1. How does the relationship between gender and age affect insomnia and sleep apnea?
2. How does the relationship between occupation and sleep duration affect sleep quality?
3. How does the relationship between heart rate and stress levels affect sleep quality?
4. How does the relationship between heart rate and stress levels affect sleep quality?
5. How does the relationship between BMI categories impact sleep quality?
6. How does the relationship between blood pressure and heart rate affect insomnia and sleep apnea?
7. How does sleep quality affect sleep disorders?
8. How to build analytical models to determine the occurrence of sleep disorders such as insomnia and sleep apnea?

CHAPTER 3: PROJECT PLAN

3.1 PROJECT DETAILS

3.1.1. Date of the project plan

9th August 2023

3.1.2. Project vision/objective

1. How does the relationship between gender and age affect insomnia and sleep apnea?
2. How does the relationship between occupation and sleep duration affect sleep quality?
3. How does the relationship between heart rate and stress levels affect sleep quality?
4. How does the relationship between heart rate and stress levels affect sleep quality?
5. How does the relationship between BMI categories impact sleep quality?
6. How does the relationship between blood pressure and heart rate affect insomnia and sleep apnea?
7. How does sleep quality affect sleep disorders?
8. sHow to build analytical models to determine the occurrence of sleep disorders such as insomnia and sleep apnea?

3.1.3. Scope

"Sleep Health and Lifestyle Dataset" is a data set on lifestyle and sleep health taken from K Hospital. Specifically, the data set emphasizes 4 aspects: Comprehensive data on sleep (i), factors Lifestyle factors (ii), cardiovascular health (iii), sleep disorder analysis (iv).

Location: K Hospital

Number of observations: 374

Time: from August to December 2020

3.2 KEY STEPS

Task	09/8	13/8	15/8	27/8	01/9	07/9	14/9	20/9
1. Requirement gathering								
1.1. Problem definition								

Task	**09/8**	**13/8**	**15/8**	**27/8**	**01/9**	**07/9**	**14/9**	**20/9**
1.2. Requirement Specification								
1.3. Project details								
2. Database								
3. Build and run the model								
3.1. Data visualization								
Descriptive analysis								
3.2. Build model based on input datasets								
4. Develop test models								
5. Test (quality plan)								
6. Draft report								
7. Final report								

Table 1. Key steps

3.2.1. QUALITY PLAN

1. Review Activities (Review meeting participants, frequency, and so on)

2. Testing Activities (Final Test)

Step 1: Run the chosen test models.

Step 2: Select the model with the highest accuracy

Step 3: Re-test to the selected model

Step 4: Start to write the draft report

3. Backup and Recovery Strategies (In case of disk crashes, network failures, and so on)

In case the mobile device is cracked, the backup plan is to save everything on drive as soon as completing the task.

CHAPTER 4: TASK SHEET

1. Project Title: Sleep health and lifestyle
2. Activity Plan Prepared by: Nguyen Khoa + Kim Cuong + Cam Lai
3. Project start date is: 9th August 2023
4. Project end date is: 20th September 2023

Task	09/8	13/8	15/8	27/8	01/9	07/9	14/9	20/9	Teammate	Status
1. Requirement gathering									Nguyen Khoa, Kim Cuong, Cam Lai	Done
1.1. Problem definition									Nguyen Khoa	Done
1.2. Requirement Specification									Cam Lai	Done
1.3. Project details									Kim Cuong	Done
2. Database									Nguyen Khoa, Kim Cuong, Cam Lai	Done
3. Build and run the model									Nguyen Khoa, Kim Cuong, Cam Lai	Done

Prepared By:	Date	Approved by:
Vo Nguyen Khoa , Nguyen T.Kim Cuong, Tran T.Cam Lai,	September 2023	Lam Chi Nguyen

Task	**09/8**	**13/8**	**15/8**	**27/8**	**01/9**	**07/9**	**14/9**	**20/9**	**Teammate**	**Status**
3.1. Data visualization									Nguyen Khoa, Kim Cuong, Cam Lai	Done
3.2. Descriptive analysis									Nguyen Khoa, Kim Cuong, Cam Lai	Done
3.3. Build model based on input datasets									Nguyen Khoa	Done
3.4. Develop test models									Nguyen Khoa	Done
4. Test (quality plan)									Nguyen Khoa, Kim Cuong, Cam Lai	Done
5. Draft report									Kim Cuong	Done
6. Final report									Cam Lai	Done

Table 2. Task sheet

CHAPTER 5: EXPLORATORY DATA ANALYSIS

5.1 DATA SOURCES

Name of data source	description
Sleep Health and Lifestyle Dataset	"Sleep Health and Lifestyle Dataset" is a data set on lifestyle and sleep health taken from K Hospital. Specifically, the data set emphasizes 4 aspects: Comprehensive data on sleep (i), factors Lifestyle factors (ii), cardiovascular health (iii), sleep disorder analysis (iv). Location: K Hospital Number of observations: 374 Time: from August to December 2020

5.2 DATA DESCRIPTION

5.2.1. Gender

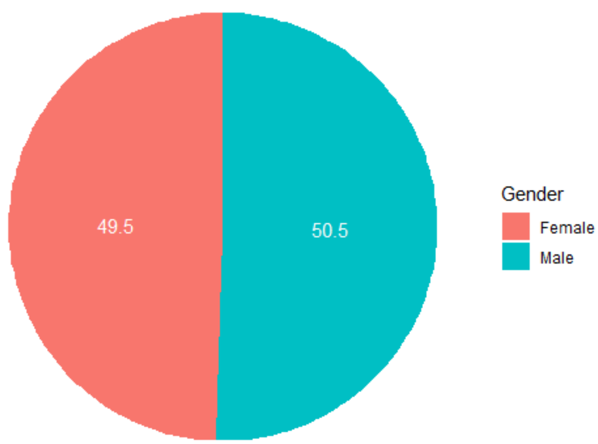

Figure 1. Gender Ratio

The gender of person. In this dataset have 185 Female(49,5%) and 189 Male(50,5%).

5.2.2. Age

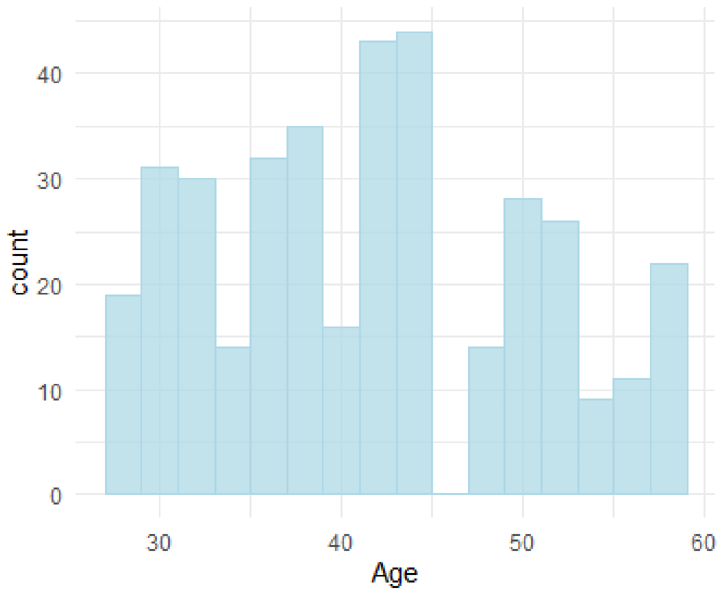

Figure 2. Age Distribution

The age of the person. In this dataset, ages are distributed between 27 – 59 year old.

5.2.3. Occupation

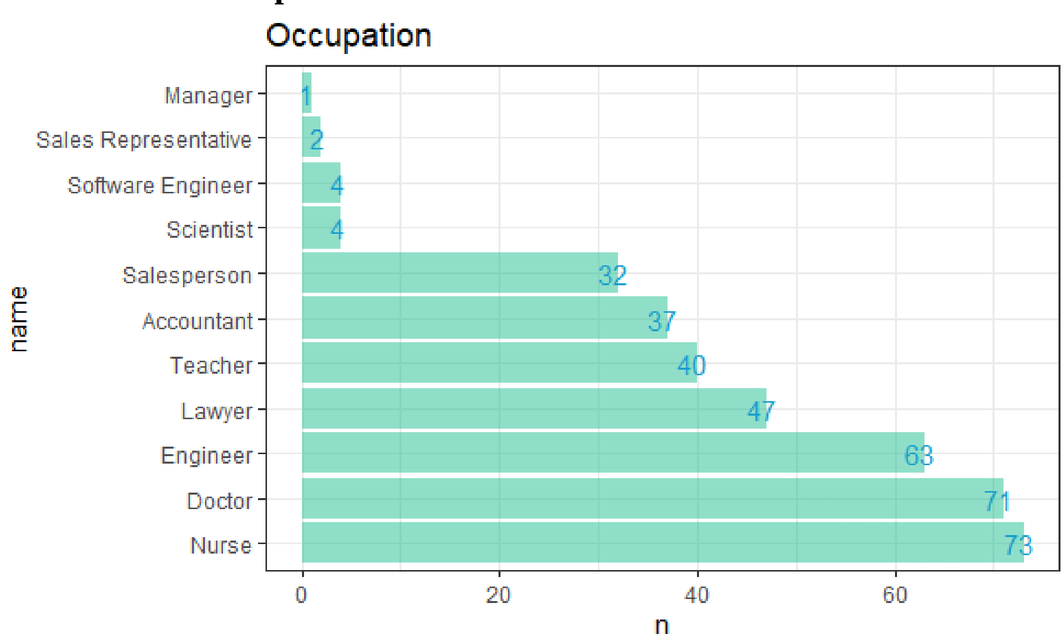

Figure 3. Count Occupation

There are 11 different vocations in this dataset. However, they are unbalanced, unequally distributed, and oriented towards professions like those of physicians, nurses, engineers,…

5.2.4. Sleep Duration

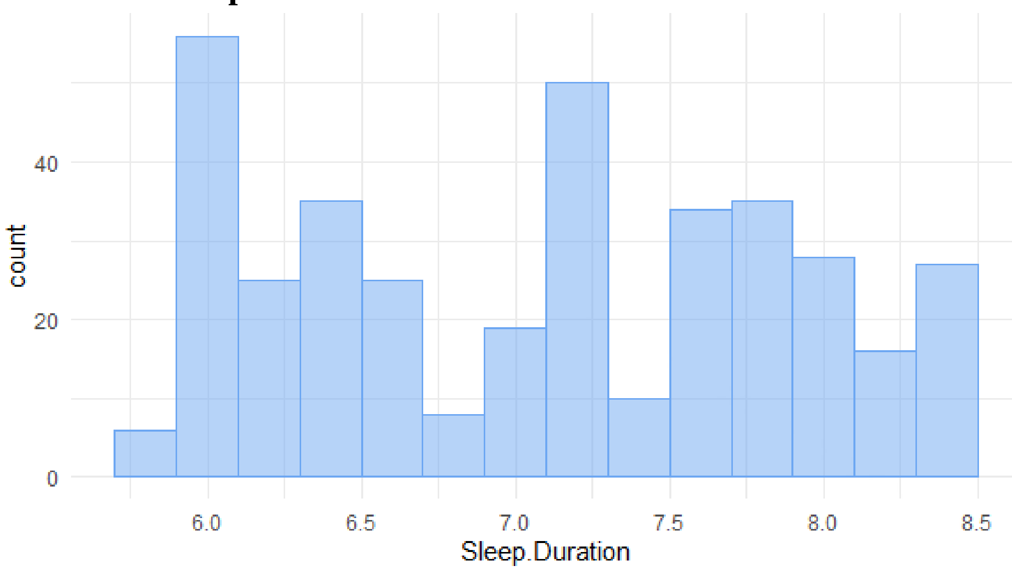

Figure 4. . Sleep Duration Distribution

The quantity of time that a person sleeps. Sleep duration may be measured for just one sleep period or over the course of a 24-hour day. his dataset's distribution of sleep duration falls between 5.8 and 8.5.

5.2.5. Quality of sleep

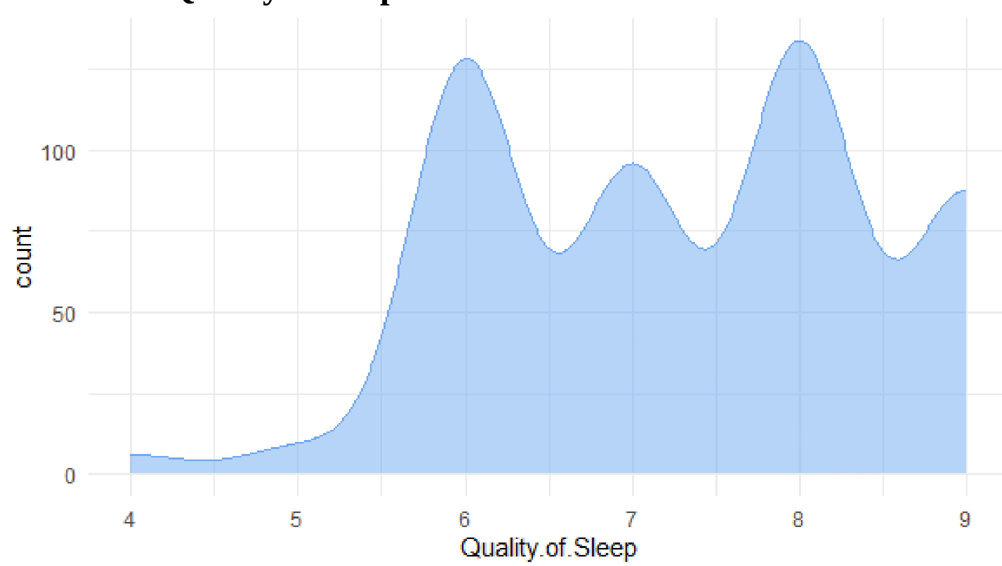

Figure 5. Quality of Sleep Distribution

A subjective rating of the quality of sleep, ranging from 1 to 10. In this dataset, Sleep quality is mainly distributed at levels 6 to 8.

5.2.6. Physical Activity Level

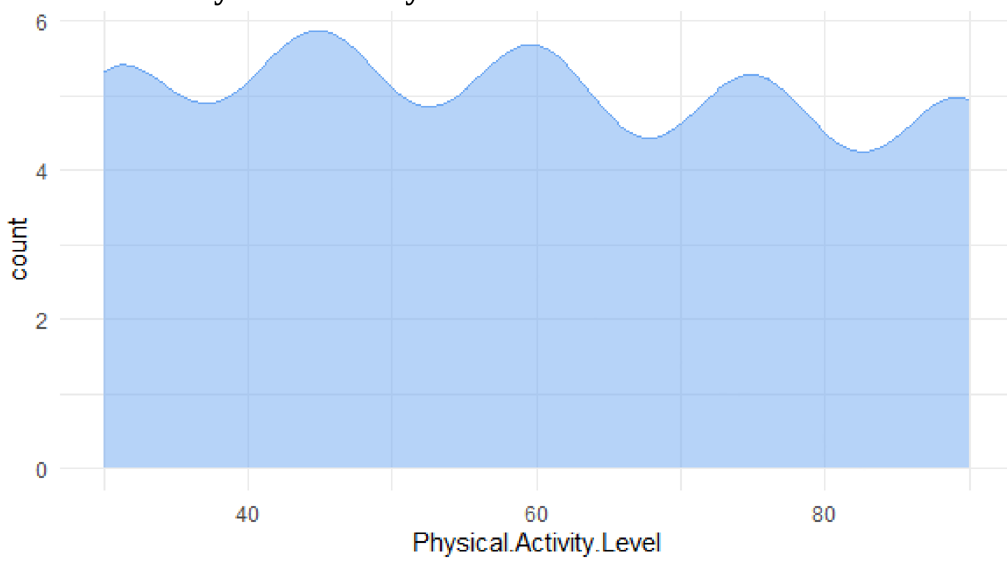

Figure 6. Physical Activity Level Distribution

The number of minutes the person engages in physical activity daily. The range of Physical Activity Levels is 30 to 90.

5.2.7. Stress Level

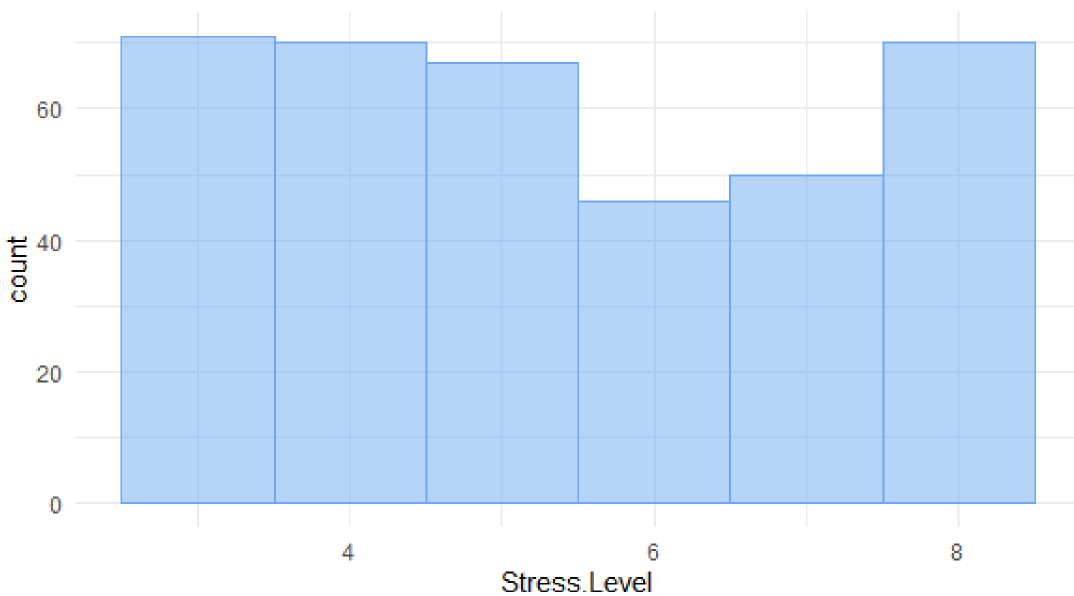

Figure 7. Stress Level Distribution

A subjective scale from 1 to 10 that represents the person's level of stress. It ranges from level 3 to 8 in this dataset..

5.2.8. BMI Category

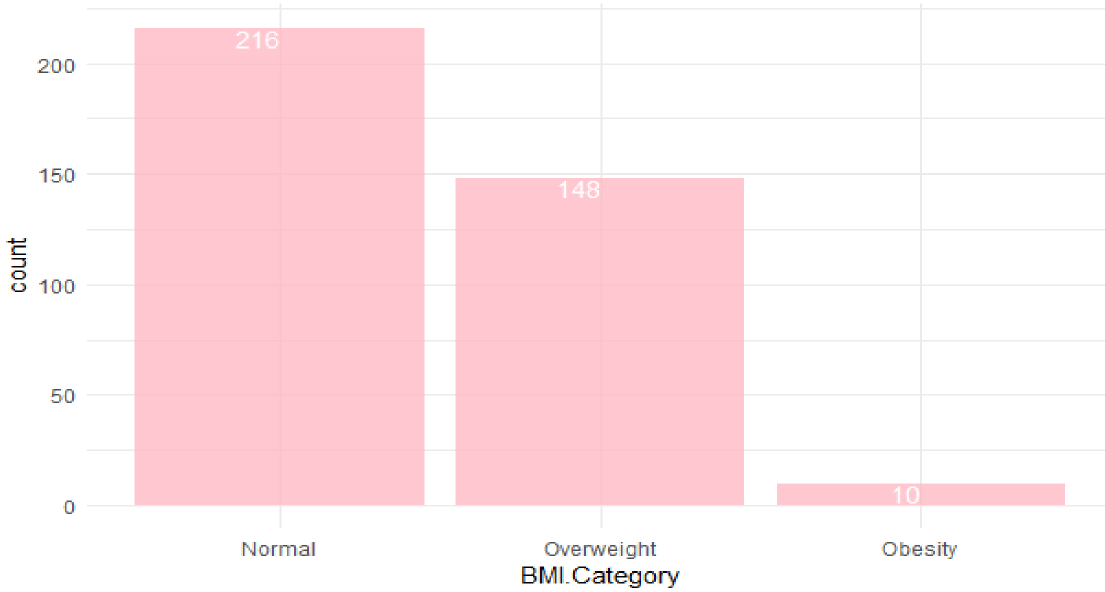

Figure 8. Count BMI

The body mass index (BMI), a measure used in medical screening, computes the ratio of your height to weight to estimate how much body fat you have. BMI is calculated by multiplying a person's height in meters (m2) by their square of weight in kilograms (kg).

The BMI ranges (in kg/m^2) classify different types:

- Underweight: Less than 18.5
- Normal: 18.5 to 24.9
- Overweight: 25 to 29.9
- Class I obesity: 30 to 34.9.
- Class II obesity: 35 to 39.9.
- Class III obesity: More than 40.

In this dataset, the BMI category of the person is divided into three groups:

- Underweight
- Normal
- Overweight.

5.2.9. Blood Pressure

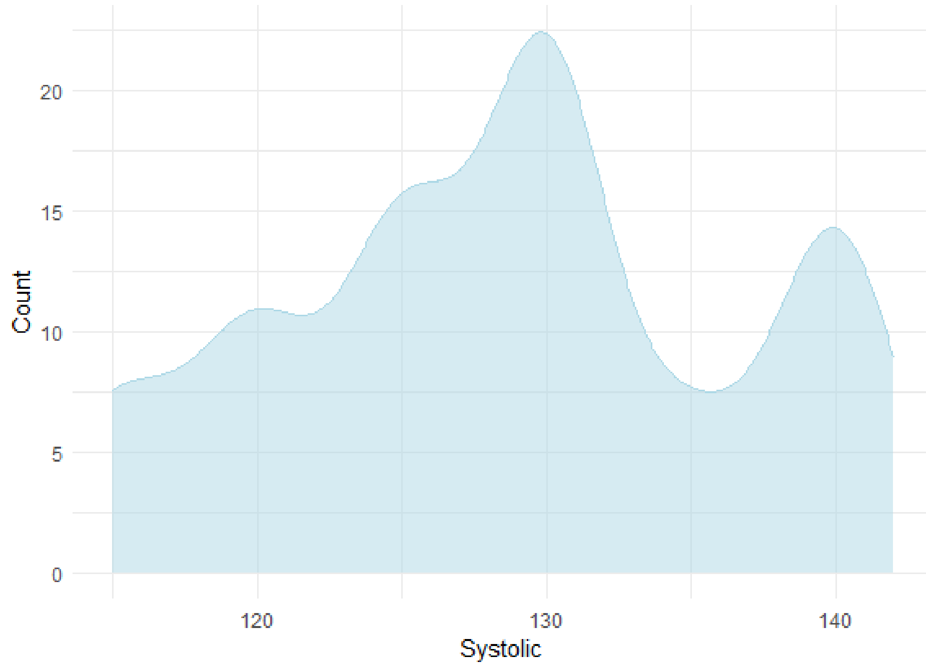

Figure 9. Systolic Distribution

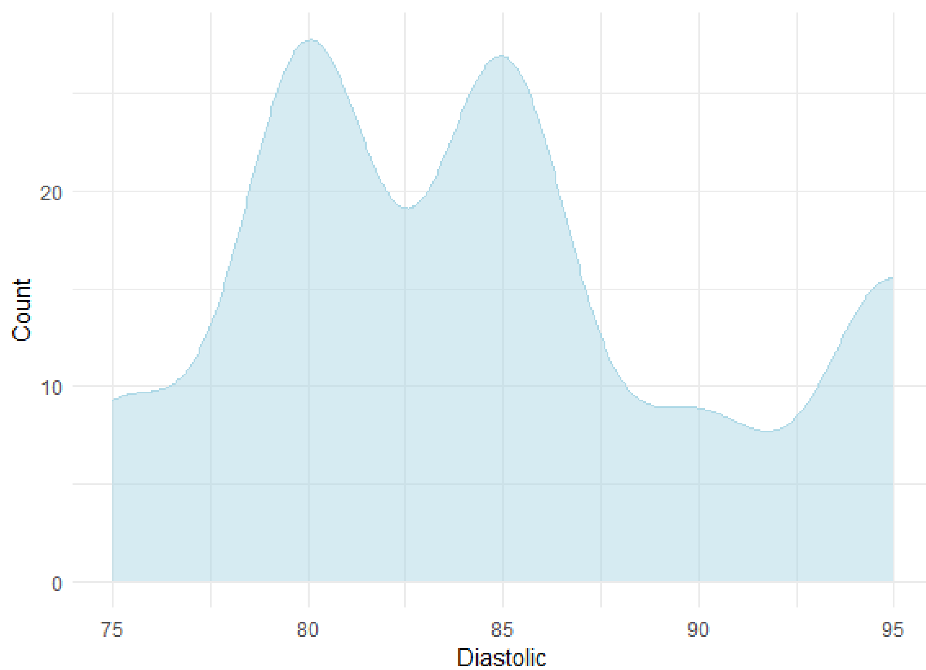

Figure 10. Diasotlic Distribution

Blood pressure is a measure of the force that your heart uses to pump blood around your body. The blood pressure is measured in millimetres of mercury (mmHg) and is given as 2 figures:

- Systolic pressure – the pressure when your heart pushes blood out
- Diastolic pressure – the pressure when your heart rests between beats

5.2.10. Heart Rate

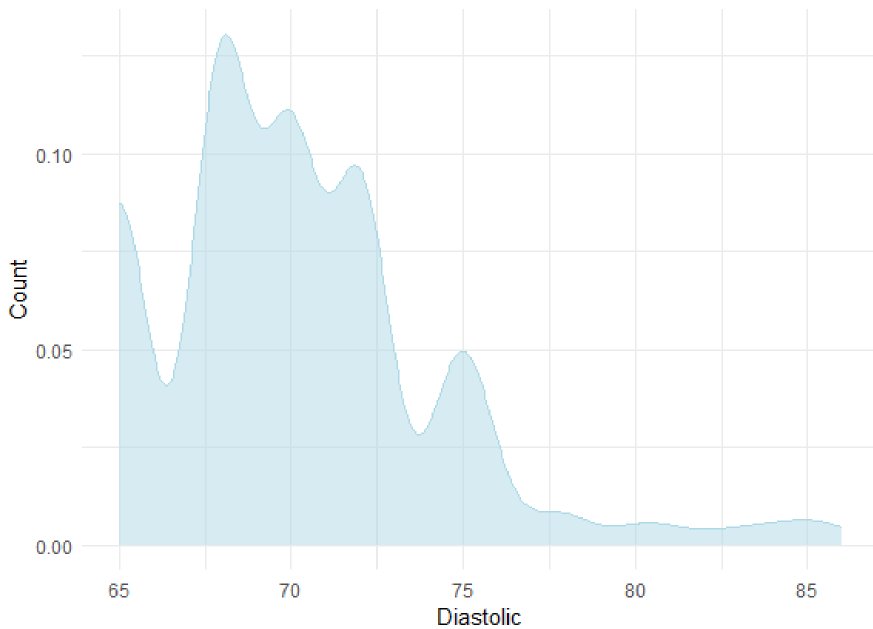

Figure 11. Heart Rate Distribution

The resting heart rate of the person in beats per minute. It ranges from approx 68-86.

5.2.11. Daily Steps

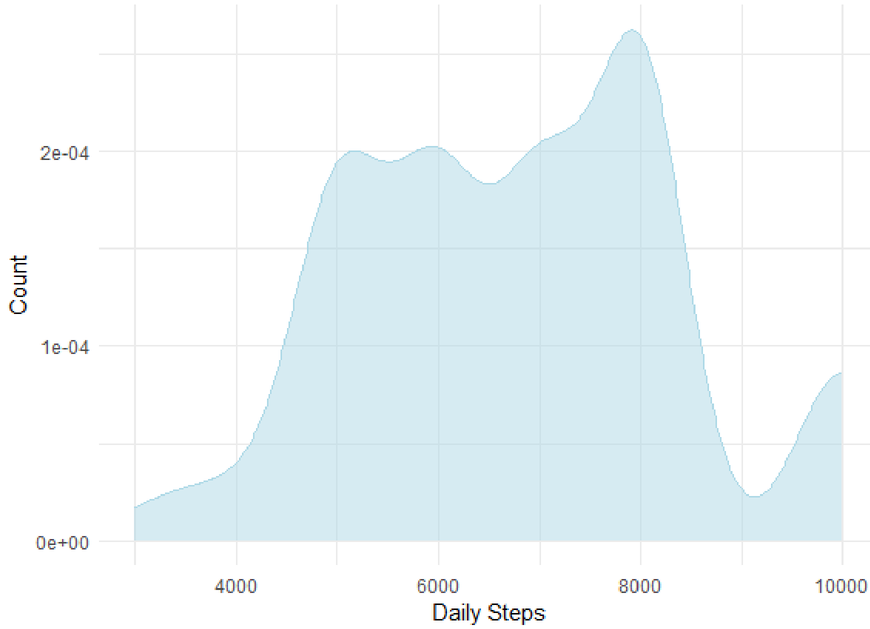

Figure 12. Daily Steps Distribution

The number of steps the person takes per day. It ranges from approx 3000 to 10000

5.2.12. Sleep Disorder

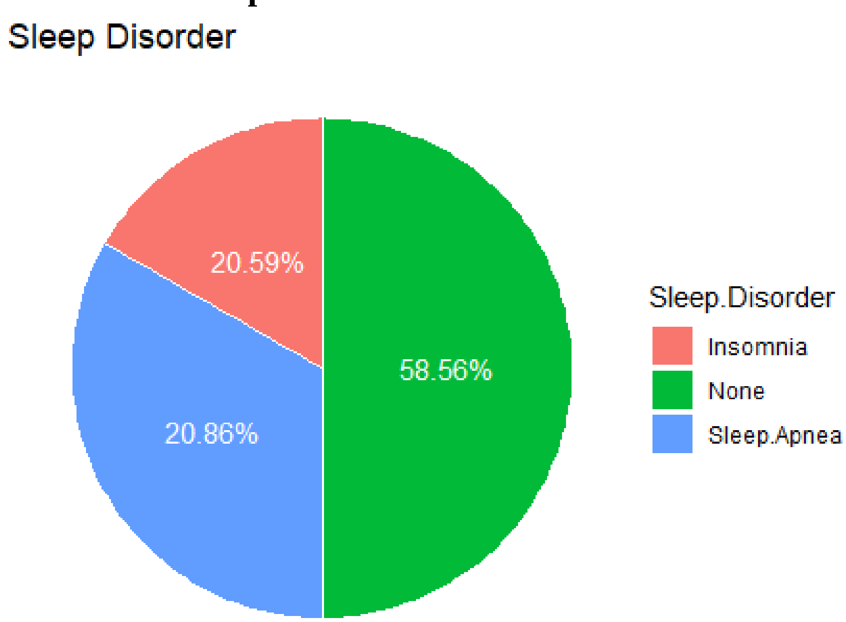

Figure 13. Sleep Disorder Ratio

Sleep disorders are conditions that result in changes in the way that you sleep. Your general health, safety, and enjoyment of life may be impacted by a sleep disturbance. Lack of sleep can make it more difficult for you to drive safely and raise your chance of developing other health issues. There are several varieties of sleep disorders. Another way to categorize sleep disorders is by habits, issues with your normal sleep-wake cycles, respiratory issues, difficulties falling asleep, and how drowsy you feel during the day.]

Some common types of sleep disorders include:

- Insomnia, a condition in which you have trouble getting asleep or remaining asleep all night.
- Sleep apnea, which causes irregular breathing patterns while you're asleep. Sleep apnea may take many different forms.
- Restless legs syndrome (RLS), a type of sleep movement disorder. When you try to fall asleep, restless legs syndrome, also known as Willis-Ekbom illness, creates an uncomfortable feeling and the impulse to move your legs.
- Narcolepsy, a condition characterized by extreme sleepiness during the day and falling asleep suddenly during the day.

In this dataset, they are divided into three groups: None, Insomnia, and Sleep Apnea. In which:

- None has 219 people
- Insomnia has 77 people
- Sleep Apnea has 78 people.

5.3 VISUALIZATION AND BASICS EVALUATION (DESCRIPTIVE ANALYTICS)

5.3.1. The relationship between gender and age affect insomnia and sleep apnea?

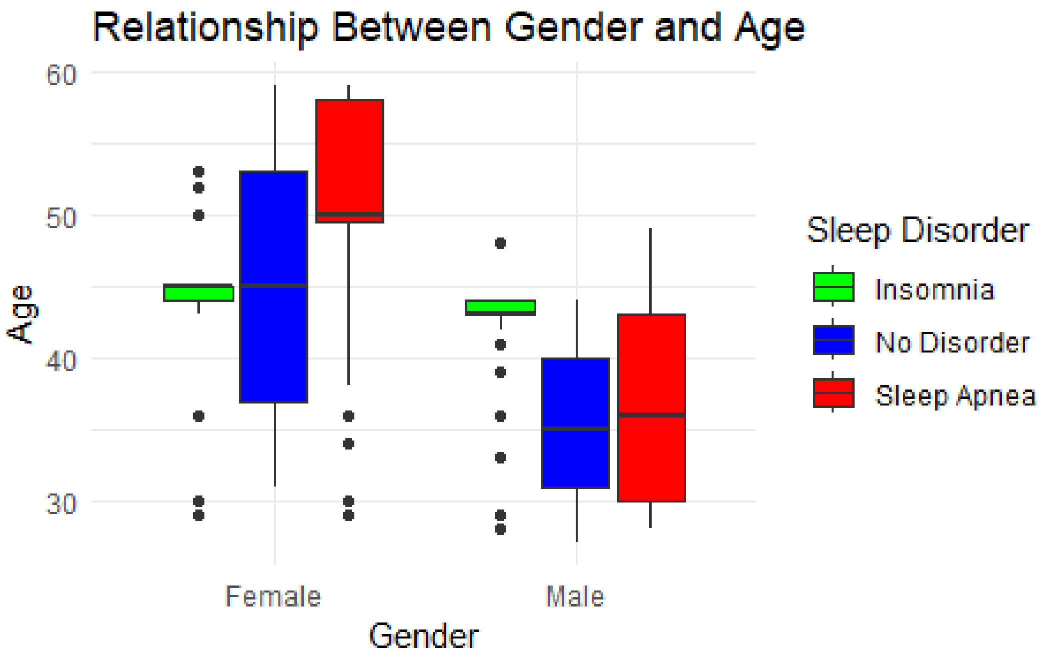

Figure 14. Relationship Between Gender and Age

- Based on the chart we can suspect that age can affect the state of insomnia ("Insomnia") and sleep apnea ("Sleep Apnea").
- In general, the ages of the Female group and the Male group have a statistically significant difference.
- In the case of Sleep Apnea, the age difference between the two groups was statistically significant, and the Female group had a higher average age than the Male group.

5.3.2. The relationship between occupation and sleep duration affect sleep quality?

To test the correlation between occupation and sleep duration that affects sleep quality, we can use statistical methods such as ANOVA analysis.

```
Df Sum Sq Mean Sq F value Pr(>F)
Occupation                10 241.91   24.19  195.50 <2e-16 ***
Sleep.Duration             1 223.35  223.35 1805.01 <2e-16 ***
Occupation:Sleep.Duration   8  25.33    3.17   25.59 <2e-16 ***
Residuals                354  43.80    0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Based on the results of ANOVA analysis, we can draw the following important points about the correlation between occupation, sleep duration and sleep quality:
- Occupation affects sleep quality: The F value of occupation (Occupation) is very high (195.50) and the p-value is low (close to 0). This suggests that there are significant differences in sleep quality between different occupations.
- Sleep duration affects sleep quality: The F value for sleep duration is also very high (1805.01) and the p-value is close to 0. This shows that sleep duration is correlated. strong with sleep quality. This means that people with different sleep durations have different sleep quality.
- The interaction between occupation and sleep duration is also important: The interaction between occupation and sleep duration (Occupation:Sleep Duration) also has a high F value (25.59) and p-value close to 0. This shows that found that the interaction between occupation and sleep duration also plays an important role in influencing sleep quality.
- In summary, from these results we can conclude that both occupation and sleep duration have a significant impact on sleep quality. In addition, the interaction between occupation and sleep duration also plays an important role in affecting sleep quality. This can provide valuable information on how to improve sleep quality in work environments and everyday life.

5.3.3. The relationship between Stress levels affect Heart Rate

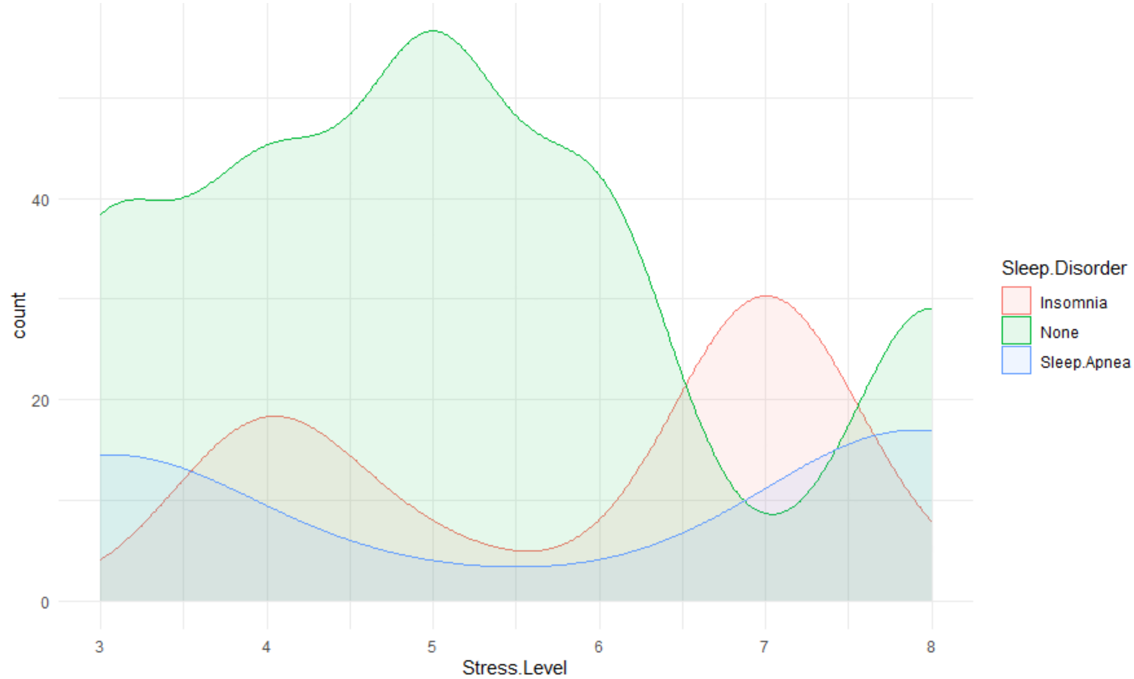

Figure 15. Distribution of Stress Level across Sleep Disorder

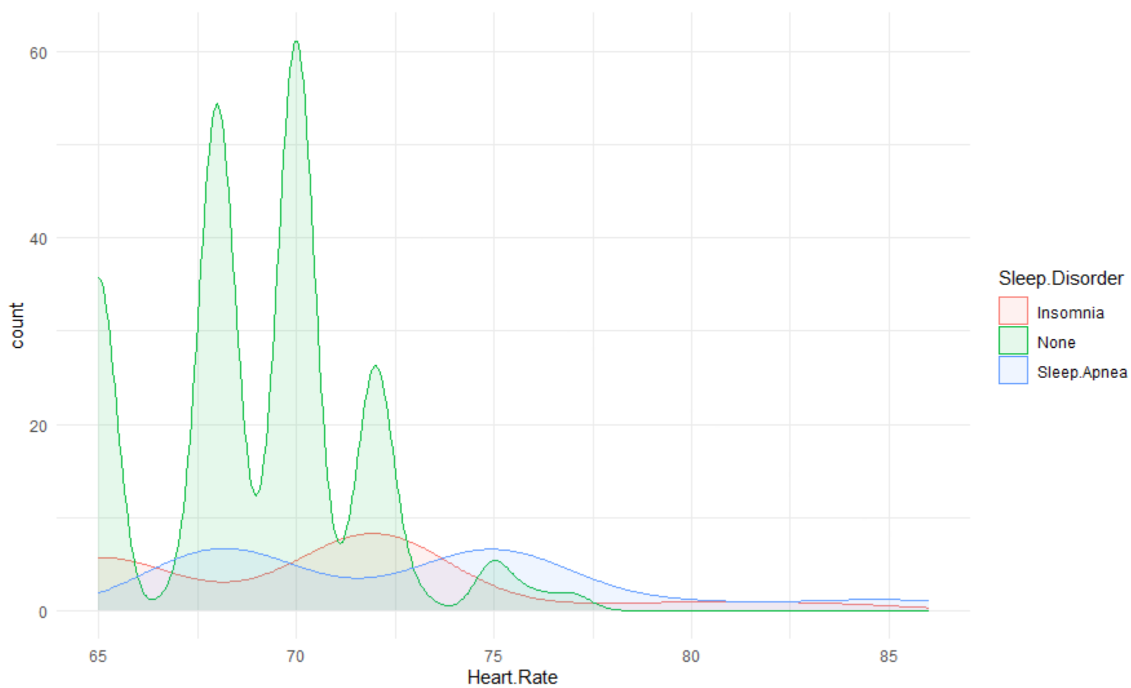

Figure 16. Distribution of Heart Rate across Sleep Disorder

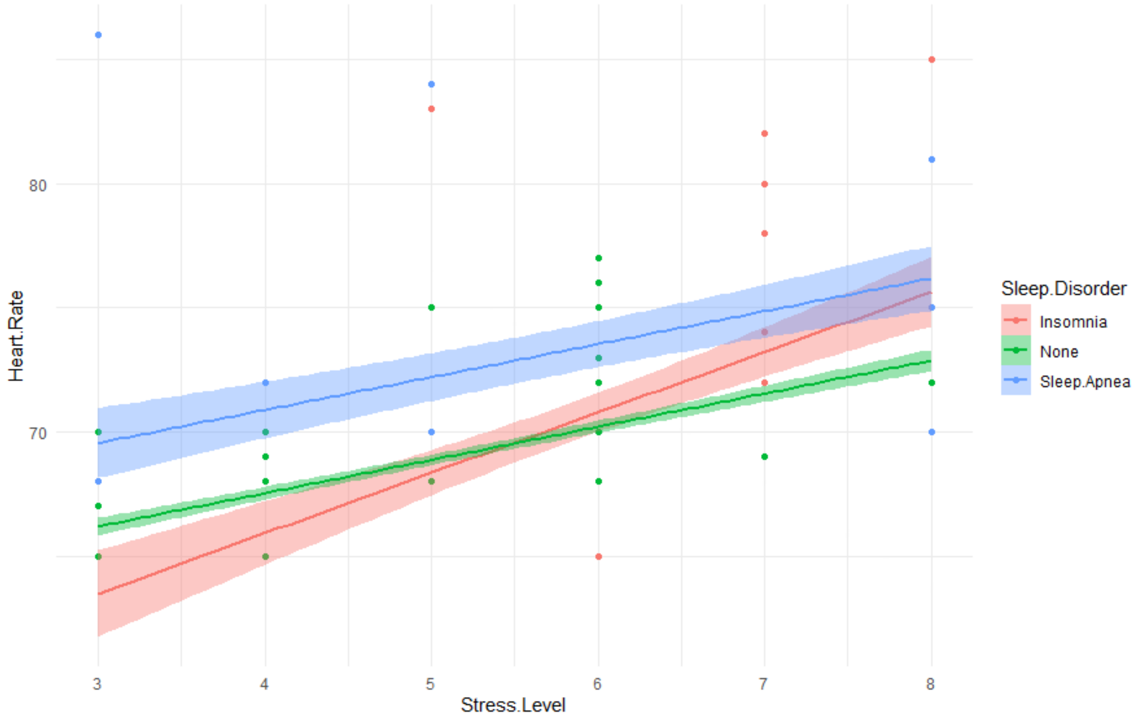

Figure 17. Relationship between Stress Level, Heart Rate and Sleep Disorder

An average stress level for healthy individuals is less than 6, and their heart rates vary from 65 to 77 beats per minute. In contrast, persons who struggle with insomnia frequently report stress levels between 7 and 8, as well as resting heart rates between 72 and 85 beats per minute. As a result, those who experience significant amounts of stress may also experience sleeplessness.

Figures 15 and 16 show that persons with sleep apnea have stress levels that are fairly uniformly distributed from 3 to 8, and that their heart rates are primarily in the range of 65 to 80. proving that the illness is not greatly influenced by variables like stress and heart rate.

5.3.4. The relationship between physical activity and daily steps impact Quality of Sleep and Sleep Disorder

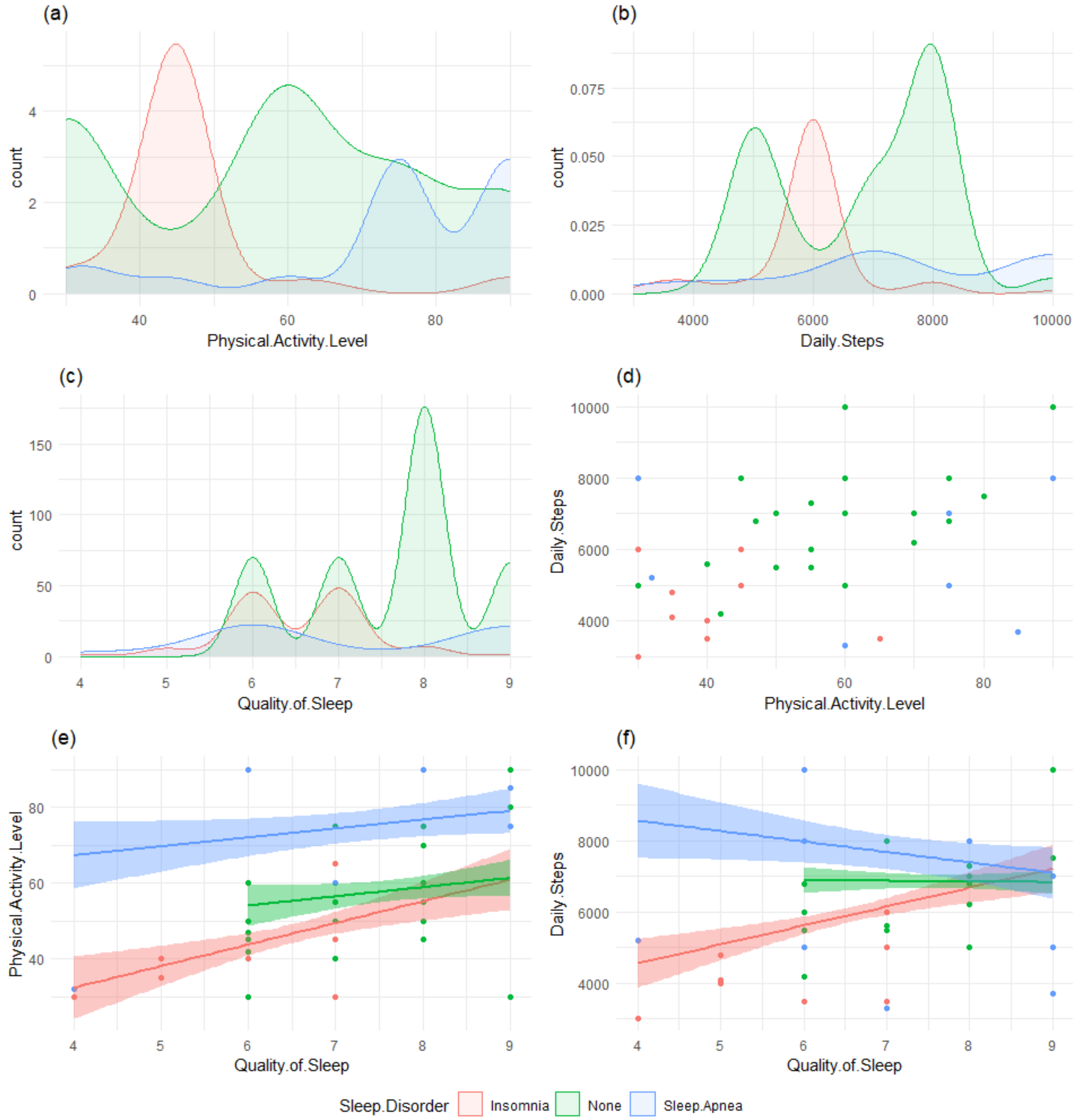

Figure 18. The relationship between Physical Activity, Daily Steps impact Quality of Sleep and Sleep Disorder

In Figure 21:

- (a) – Distribution of Physical Activity Level across Sleep Disorder
- (b) – Distribution of Daily Steps across Sleep Disorder

- (c) – Distribution of Quality of Sleep across Sleep Disorder
- (d) – Relationship between Daily Steps, Physical Activity and Sleep Disorder
- (e) – Relationship between Quality of Sleep, Physical Activity and Sleep Disorder
- (f) – Relationship between Quality of Sleep, Daily Steps and Sleep Disorder

Through the chart of Physical Activity level and Daily Steps, the correlation between these two indicators can be seen:

- People without the disease have a daily step count of 7000 or more, a high level of physical activity of higher than 60, and a sleep quality of this group of 6 or more. From there, it is clear that increasing physical activity can improve sleep quality and decrease sleep problems.
- People who have insomnia engage in physical activity at a level between 35 and 55, take 5500 to 6500 steps per day, and have sleep of a quality between 4 and 6. demonstrating that in this group of persons, a lack of physical exercise had an influence on insomnia.
- People with sleep apnea have a physical activity level of 70 or higher do between 6000 and 9000 steps per day, and report sleeping between 5 and 9. This demonstrates that in this population, physical activity had no impact on sleep apnea. However, physical activity still affects sleep quality and BMI of this group of people.

5.3.5. The relationship between BMI categories affecting sleep quality

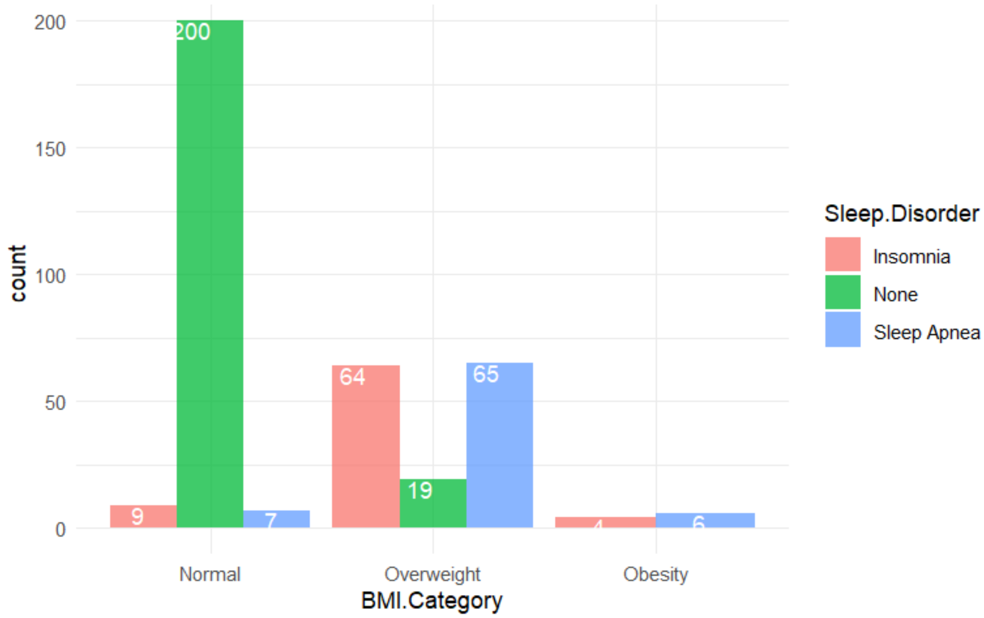

Figure 19. . Count BMI Categories

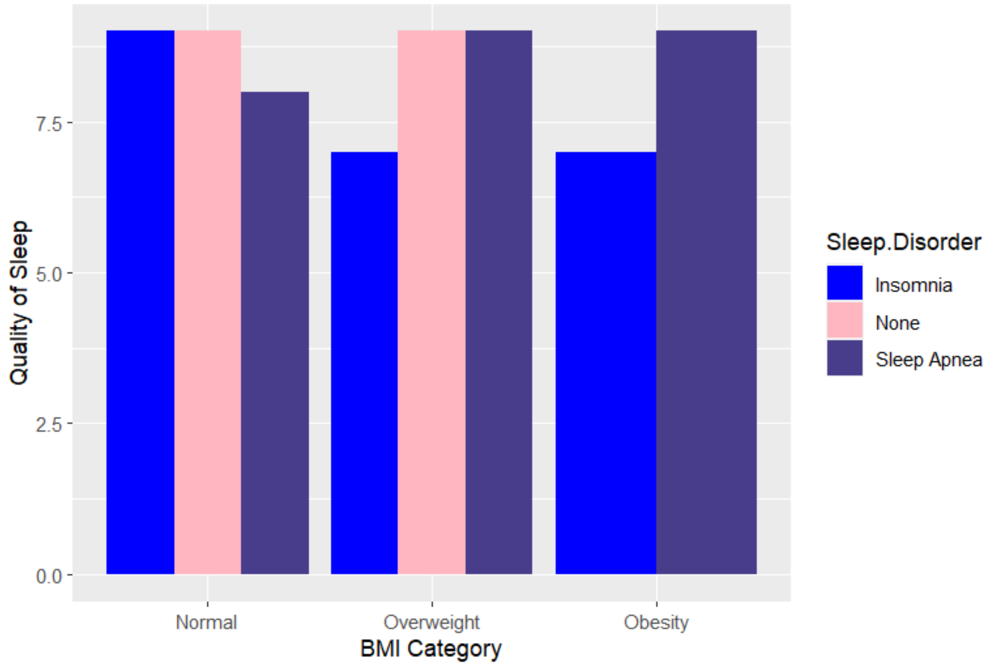

Figure 20.Relationship between BMI Category, Quality of Sleep and Sleep Disorder

Correlation chart between BMI categories affecting sleep quality of 3 groups of subjects: normal weight, overweight, and obesity.

- For the group of subjects with normal weight, sleep quality is relatively high, most of them will sleep well and their sleep quality is close to the data limit of 8.5. However, there are also a small number of people who have normal weight but suffer from insomnia and sleep apnea. But for people with insomnia, most of their sleep quality was close to the data limit of 8.5. As for subjects with sleep apnea, their sleep quality will be lower than subjects without the disease and with insomnia and their sleep quality will fall in the range of 5.8-7.5.
- For the group of overweight subjects, the opposite is true for the group of subjects with normal weight. The sleep quality of overweight subjects with insomnia and sleep apnea is equal.
- For obese subjects, they definitely suffer from two sleeping diseases: insomnia and sleep apnea syndrome.

5.3.6. The relationship between blood pressure and heart rate affect insomnia and sleep apnea

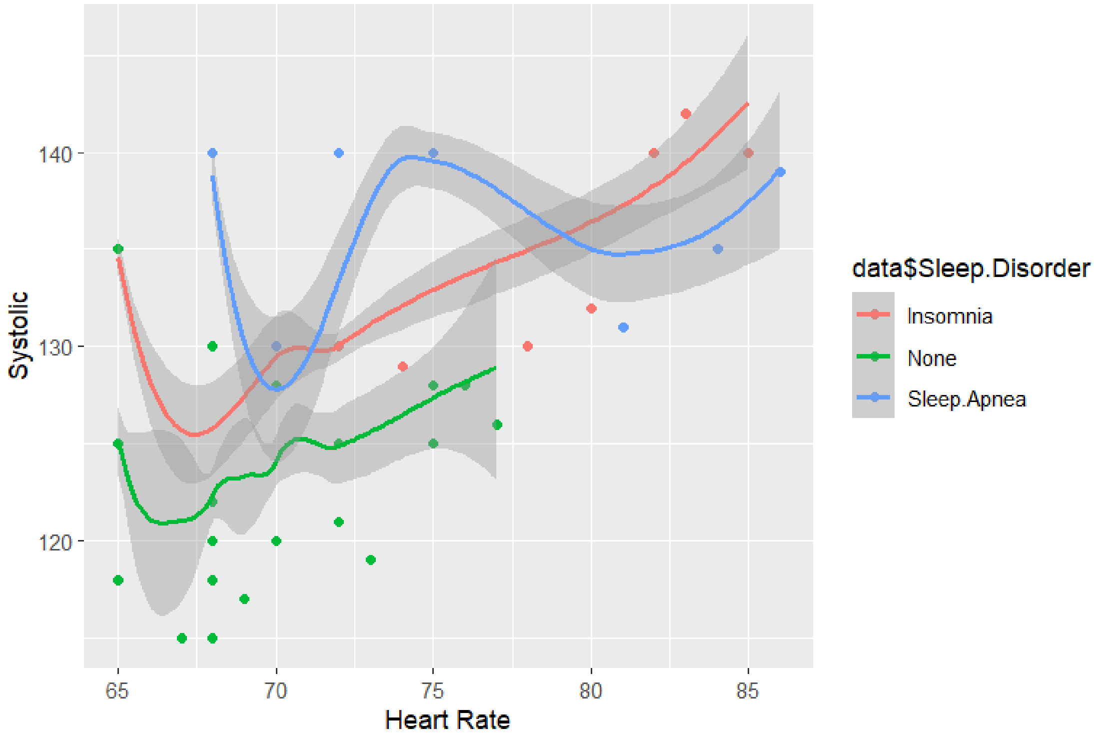

Figure 21. Relationship between Heart Rate, Systolic(Blood Pressure) and Sleep Disorder

- The figure 21 shows that for subjects without sleeping diseases (blue), systolic blood pressure fluctuates between 115 and 135 and heart rate fluctuates between 65 and 77 beats/minute. rest. Shows us that the systolic blood pressure and heart rate of the group of subjects without sleeping diseases are in a stable range.
- For subjects with insomnia, systolic blood pressure fluctuates between 130 and 142 and heart rate ranges from 72 to 85 beats/minute at rest. Through the expression of systolic blood pressure and heart rate, we can see that they belong to the high blood pressure group, because systolic blood pressure skyrocketed from a normal state of 130 to 142. Showing us that high systolic blood pressure (high blood pressure) will lead to insomnia. And according to medical experts, blood pressure, heart rate and insomnia are closely related to each other, because blood pressure is the pressure of blood in the blood vessel walls, affected by the squeezing force of the heart. When awake and active, the heart and blood vessels have to work harder to carry blood throughout the body's organ systems and return blood to the heart. When sleeping, the heart rests more and is less active during other activities. or rest completely. Thus, if you have little sleep, lack of sleep, light sleep or insomnia for a long time, regularly, the heart and vascular system

have to work harder, easily causing disorders leading to hypertension and cardiovascular disease. Existing high blood pressure, cardiovascular disease, and prolonged insomnia will aggravate the disease and even cause complications.

- Sleep apnea syndrome is a period of cessation of breathing for 10-30 seconds during sleep, which can last for more than 1 minute, causing slow heart rate and low blood pressure. The body will now react to the sudden decrease in oxygen saturation with short awakenings so that the patient can breathe again, helping to speed up the heart rate and increase blood pressure . Looking at the chart, we see that the systolic blood pressure of the group of subjects with sleep apnea fluctuates up and down, at first it was 140, but with a heart rate of 70, the systolic blood pressure suddenly dropped to about 128 to 130 and then increase it to 140 and lower it slightly. And this also causes the patient's blood pressure to increase significantly to compensate. In the long term, blood pressure can remain high even when the patient is awake. At this time, high blood pressure will be a complication of sleep apnea syndrome.

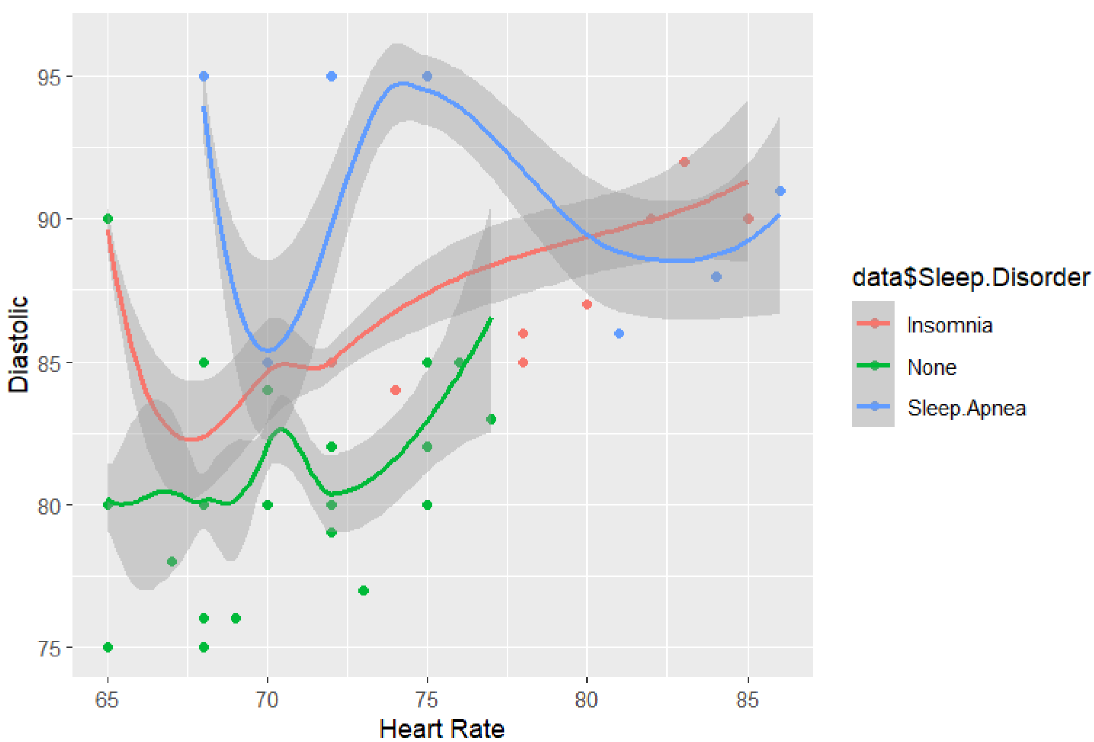

Figure 22. Relationship between Heart Rate, Díatolic(Blood Pressure) and Sleep Disorder

The figure 22 shows the relationship between diastolic blood pressure and heart rate affecting sleep disorders.

- For subjects without sleep syndromes, diastolic blood pressure fluctuates between 75 and 90 and we see it gradually increasing. This shows that people who do not suffer from sleep syndromes such as insomnia and sleep apnea have diastolic blood pressure within normal levels.
- For subjects with insomnia, diastolic blood pressure ranges from 85 to 92. According to medical experts, if diastolic blood pressure is 92 or higher, the patient will have high blood pressure. According to the chart, we see that the blood pressure index of 90 or higher in the insomnia group is not too high. We can predict that subjects with insomnia may have high blood pressure.
- For subjects with sleep apnea, high diastolic blood pressure ranges from 90 to 95 and it increases and decreases suddenly. Sleep apnea syndrome is a period of cessation of breathing for 10-30 seconds during sleep, which can last for more than 1 minute, causing slow heart rate and low blood pressure. The body will now react to the sudden decrease in oxygen saturation with short awakenings so that the patient can breathe again, helping to speed up the heart rate and increase blood pressure .

CHAPTER 6: AI/ML MODEL SPECIFICATION

6.1 MODEL NAME

Sleep Health and Lifestyle

6.2 MODEL SPECIFICATION

- Input: Dataset on lifestyle and sleep health taken from K Hospital. This gathere 374 obseration from August to December 2020. Specifically, the dataset emphasizes 4 aspects: Comprehensive data on sleep (i), factors Lifestyle factors (ii), cardiovascular health (iii), sleep disorder analysis (iv). It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

- Output: Classification of Sleep Disorder (None, Sleep Apnea, Insomnia)

- Approach: Decision Tree and Random Forest Classification

6.3 DATASET USED FOR MODEL

Data preprocessing is essential before taking the next steps in the model building process.

6.3.1. HANDLE MISSING VALUE

- There are no missing value in dataset

6.3.2. DATASETS USED FOR MODELING

Field name	Description	Data Type	Range of data	Remark
Gender	The gender of the person	Character	Male, Female	
Age	The gender of the person	Int	27 – 59	
Sleep Duration	The number of hours the person sleeps per day.	Numeric	5.8 – 8.5	

Field name	Description	Data Type	Range of data	Remark
Quality of Sleep	A subjective rating of the quality of sleep, ranging from 1 to 10.	Int	4 – 9	
Physical Activity Level	The number of minutes the person engages in physical activity daily.	Int	30 – 90	
Stress Level	A subjective rating of the stress level experienced by the person	Int	1 - 10	
BMI Category	The BMI category of the person	Character	Underweight, Normal, Overweight	
Blood Pressure (Systolic / Diastolic	The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.	Numeric	115 - 142 / 75 - 95	Split to 2 column "Systolic" and "Diastolic" while build model
Heart Rate	The resting heart rate of the person in beats per minute.	Int	65 - 86	
Daily Steps	The number of steps the	Int	3000 – 10000	

Field name	Description	Data Type	Range of data	Remark
	person takes per day.			
Sleep Disorder	The presence or absence of a sleep disorder in the person.	Character	None, Insomnia, Sleep Apnea	

Table 3. Datasets used for modeling

6.4 PROCESS AND CODE ANALYSIS

6.4.1. Decision Tree

Decision tree algorithms use the training data to segment the predictor space into non-overlapping regions, the nodes of the tree. Each node is described by a set of rules which are then used to predict new responses. The predicted value for each node is the most common response in the node (classification), or mean response in the node (regression).

The algorithm splits by recursive partitioning, starting with all the observations in a single node. It splits this node at the best predictor variable and best cutpoint so that the responses within each subtree are as homogenous as possible, and repeats the splitting process for each of the child nodes until a stopping criterion is satisfied.

This much results in a large tree that provides a good fit to the training data, but it likely overfits the data. The solution is to “prune” leaves from the tree. The most common pruning method is cost-complexity pruning. Cost-complexity pruning minimizes the cost complexity:

$$CC(T) = R(T) + cp|T|$$

Where $|T|$ is the size of tree (complexity), $R(T)$ is the missclassification rate (decision trees) or RSS (regression trees), and $cp|T|$ is the complexity parameter.

It is expensive to evaluate the error on all possible subtrees, so instead the algorithm defines a sequence of nested trees by successively pruning leaves from the tree, repeating until only the root node remains. The complexity parameter yielding the lowest cost complexity is the optimal tree size.

Step 1: Set worrking enviroment and import data

```
setwd("H:/My Drive/R Project/SleepHealthandLifestyle")
data <- read.csv("Sleep_health_and_lifestyle_dataset.csv")

data <- data[, c(2:3, 5:13)]
```

Step 2: Check the data class and transform data

```
data$Blood.Pressure <- sub("/", " ", data$Blood.Pressure)
data <- separate(data, Blood.Pressure, into = c("Systolic", "Diastolic"), sep = " ")
data$Systolic <- as.numeric(data$Systolic)
data$Diastolic <- as.numeric(data$Diastolic)
data$BMI.Category<- gsub("Normal Weight", "Normal", data$BMI.Category)
data$BMI.Category<- gsub("Obese", "Obesity", data$BMI.Category)
data$Sleep.Disorder[data$Sleep.Disorder == "Sleep Apnea"] = "Sleep.Apnea"
data <- data %>% transform(Sleep.Disorder = as.factor(Sleep.Disorder))
data$Gender <- as.factor(data$Gender)
data$BMI.Category <- as.factor(data$BMI.Category)
```

Step 3: Divide data into training and testing group and run the demo test

```
set.seed(111)
ind <- sample(2, nrow(data), replace = TRUE, prob = c(0.8, 0.2))
train_data <- data[ind==1,]
test_data <- data[ind==2,]

addmargins(prop.table(table(ind)))
```

- Using Rpart library for build Decision tree model

```
decision_tree<- rpart(Sleep.Disorder ~ .,
                      data = train_data,
                      method = 'class',
                      xval = 10, # k-fold với k = 10
                      )

printcp(decision_tree)
```

```
##
## Classification tree:
## rpart(formula = Sleep.Disorder ~ ., data = train_data, method
= "class",
##     xval = 10)
##
## Variables actually used in tree construction:
## [1] Age                    BMI.Category
Physical.Activity.Level
## [4] Systolic
##
## Root node error: 125/297 = 0.42088
##
## n= 297
##
##      CP nsplit rel error xerror     xstd
## 1 0.376      0     1.000  1.000 0.068066
## 2 0.344      1     0.624  0.800 0.065155
## 3 0.040      2     0.280  0.288 0.044997
## 4 0.010      4     0.200  0.320 0.047066
```

```
rpart.plot(decision_tree, yesno = TRUE)
```

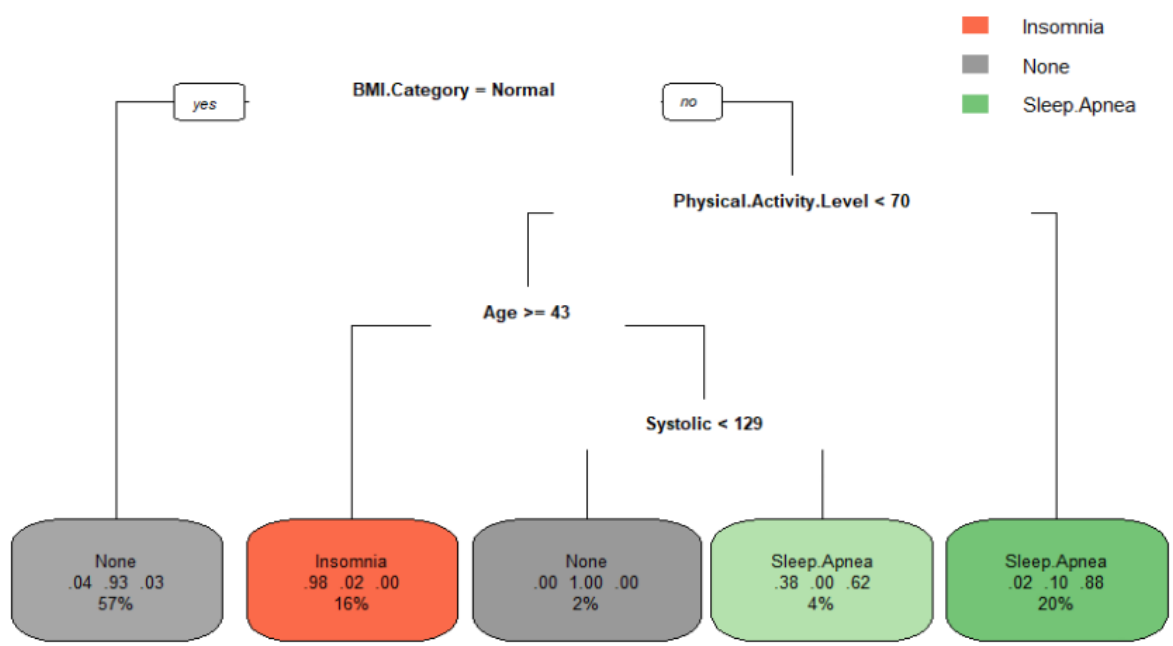

Figure 23. Decision tree

```
rpart.rules(decision_tree)
```

```
##  Sleep.Disorder  Inso None Slee
##        Insomnia [ .98  .02  .00] when BMI.Category is Obesity or Overweight & Physical.Activity.Level <  70 & Age >= 43
##            None [ .04  .93  .03] when BMI.Category is Normal
##            None [ .00 1.00  .00] when BMI.Category is Obesity or Overweight & Physical.Activity.Level <  70 & Age <  43 & Systolic <  129
##     Sleep.Apnea [ .38  .00  .62] when BMI.Category is Obesity or Overweight & Physical.Activity.Level <  70 & Age <  43 & Systolic >= 129
##     Sleep.Apnea [ .02  .10  .88] when BMI.Category is Obesity or Overweight & Physical.Activity.Level >= 70
```

The Figure 23 shows the rules after apply Decision Tree algorithm on Traning-data. For example, *"If BMI Category is Normal then Sleep Disorder is None", "If BMI Category is Obesity or Overweight and Physical Activity Level less than 70 and Age greater than or equal 43 then Sleep Disorder is Insomnia",…*

Command **rpart.rules**(decision_tree) will display rules that have the result after classification as their first word, followed by the percentage of the result in the data group and then the rule to give the result.

```
plotcp(decision_tree)
```

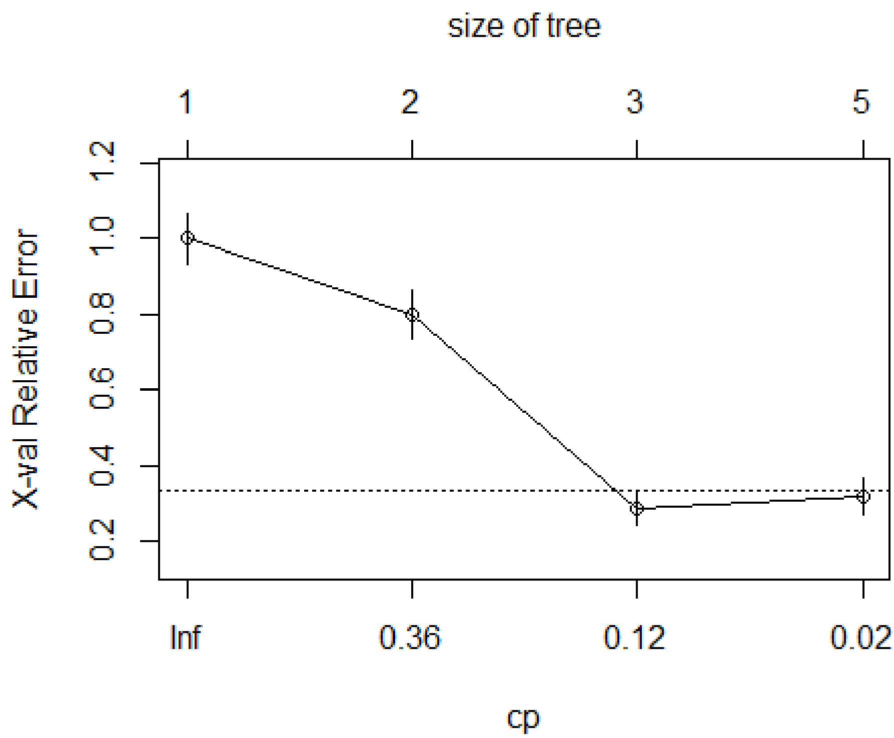

Figure 24. X-Val Relative Error

Step 4: Prune the tree to optimal size

From Figure 24, the dashed line is set the minimun **`xerror + xstd`**. Any value below the line would be considered statistically significant. A good choice for CP is often the largest value for which the error is within a standard deviation of the mimimum error.

A good way to detect and capture the correct samllest **`cp`** is with the **`which.min()`** function. Use the **`prune()`** function to prune the tree by specifuying the associated cost-complexity **`cp`**.

```
library(caret)
decision_tree_prune <- prune(decision_tree, cp = decision_tree$cptable[which.min(decision_tree$cptable[, "xerror"]), "CP"])

printcp(decision_tree_prune)
```

```
##
## Classification tree:
```

```
## rpart(formula = Sleep.Disorder ~ ., data = train_data, method
= "class",
##     xval = 10)
##
## Variables actually used in tree construction:
## [1] BMI.Category            Physical.Activity.Level
##
## Root node error: 125/297 = 0.42088
##
## n= 297
##
##      CP nsplit rel error xerror     xstd
## 1 0.376      0     1.000  1.000 0.068066
## 2 0.344      1     0.624  0.800 0.065155
## 3 0.040      2     0.280  0.288 0.044997
```

```
rpart.plot(decision_tree_prune, yesno = TRUE)
```

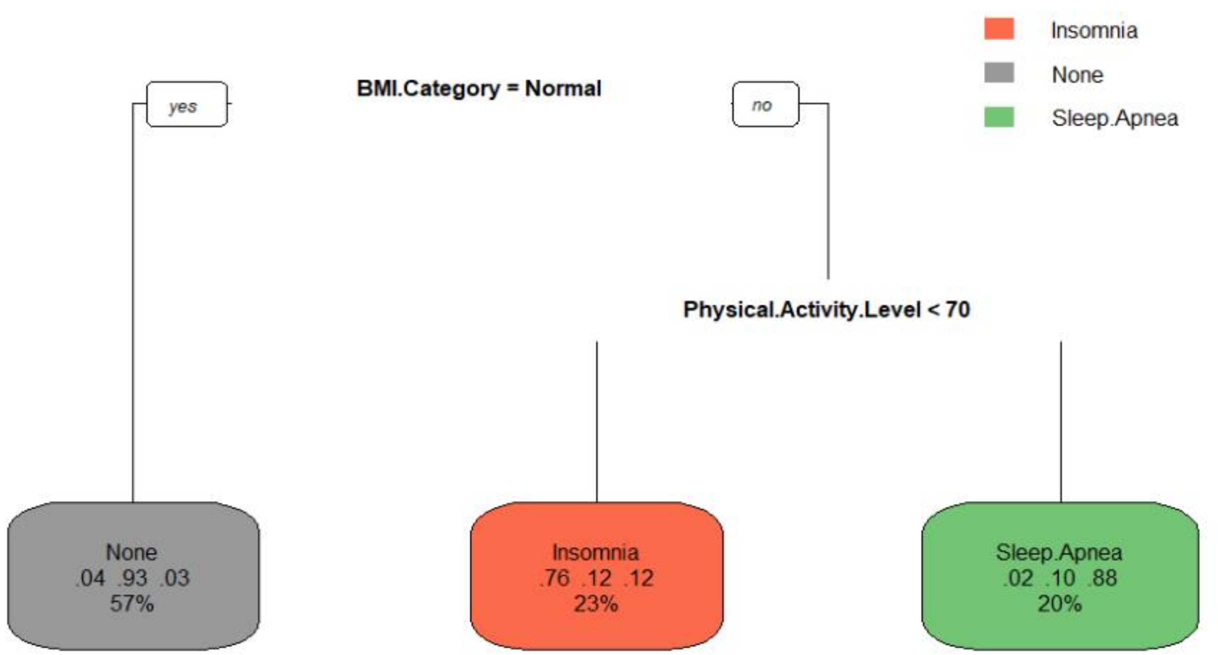

Figure 25. Decision tree after prune

Figure 25 shows that the tree has been trimmed in comparison to figure 23 and that the depth of the tree has changed. It also show that several rules in the tree before and after trimming are comparable. Some rules can be seen on Figure 25: *"If BMI Category is Normal then Sleep Disorder is None", "If BMI Category is Obesity or Overweight and Physical Activity Level less than 70 then Sleep Disorder is Insomnia",…*

```
(rpart.rules(decision_tree_prune))
```

```
##  Sleep.Disorder  Ins Non Sle
##        Insomnia [.76 .12 .12] when BMI.Category is Obesity or
Overweight & Physical.Activity.Level <  70
##            None [.04 .93 .03] when BMI.Category is
Normal
##     Sleep.Apnea [.02 .10 .88] when BMI.Category is Obesity or
Overweight & Physical.Activity.Level >= 70
```

Step 5: Save Decision tree Model

```
save(decision_tree_prune, file = "DecisionTree.rda")
```

6.4.2. Random Forest

Random forest is a commonly-use machine learning algorithm, which combines the output of multiple decision trees to reach a single result.

Before training, the three primary hyperparameters for random forest algorithms must be established. These include the size of the nodes, the number of trees, and the sample size of the features. The random forest classifier may then be used to address issues with regression or classification.

Step 1: set working environment and import data.

```
data <- read.csv("Sleep_health_and_lifestyle_dataset.csv")
data <- data[, c(2:3, 5:13)]
```

Step 2: Check data and Transform data.

```
summary(data)

data$Blood.Pressure <- sub("/", " ", data$Blood.Pressure)
data <- separate(data, Blood.Pressure, into = c("Systolic",
"Diastolic"), sep = " ")
data$Systolic <- as.numeric(data$Systolic)
data$Diastolic <- as.numeric(data$Diastolic)
data$BMI.Category<- gsub("Normal Weight", "Normal",
data$BMI.Category)
data$BMI.Category<- gsub("Obese", "Obesity", data$BMI.Category)
data$Sleep.Disorder[data$Sleep.Disorder == "Sleep Apnea"] =
"Sleep.Apnea"
data <- data %>% transform(Sleep.Disorder =
```

```
as.factor(Sleep.Disorder))
data$Gender                    <-                    as.factor(data$Gender)
data$BMI.Category <- as.factor(data$BMI.Category)
```

Step 3: Divide data into training and testing group and run the demo model

```
set.seed(111)
ind <- sample(2, nrow(data), replace = TRUE, prob = c(0.8, 0.2))
train_data                    <-                    data[ind==1,]
test_data                     <-                    data[ind==2,]

addmargins(prop.table(table(ind)))

rf_model  <-  randomForest(Sleep.Disorder~.,  data  =  train_data,
ntree        =        1000,        important        =        TRUE)
rf_model
```

```
##
## Call:
##  randomForest(formula = Sleep.Disorder ~ ., data = train_data,
ntree = 1000, important = TRUE)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 8.75%
## Confusion matrix:
##             Insomnia None Sleep.Apnea class.error
## Insomnia          47    5           7  0.20338983
## None               1  166           5  0.03488372
## Sleep.Apnea        5    3          58  0.12121212
```

```
oob_err_data                    <-                    data.frame(
  Trees        =        rep(1:nrow(rf_model$err.rate),        4),
  Type = rep(c("OOB", "Insomnia", "None", "Sleep.Apnea"), each =
nrow(rf_model$err.rate)),
  Error                 =                c(rf_model$err.rate[,"OOB"],
rf_model$err.rate[,"Insomnia"],       rf_model$err.rate[,"None"],
rf_model$err.rate[,                              "Sleep.Apnea"]))
```

```
ggplot(data = oob_err_data, aes(x = Trees, y= Error)) +
geom_line(aes(color = Type))+ theme_minimal()
```

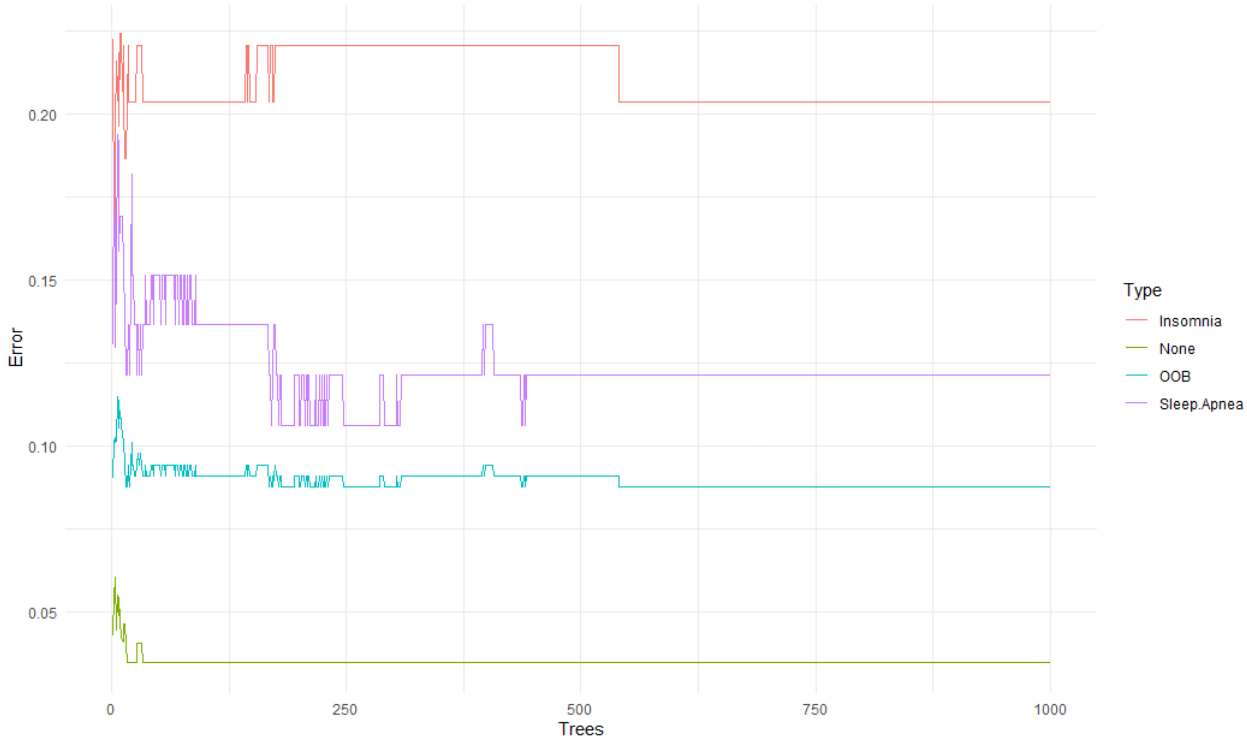

Figure 26. Random forerst Error

Form Figure 26, the error of all classifiers is rather high in the range of 0 to 250 trees, more steady from 250 to 500 trees, and typically less volatile beyond 500 trees.

Step 4: Find the optimal mtry value

```
rf_fit <- train(Sleep.Disorder~.,
                  data = train_data,
                  method = "ranger",
                  tuneLength = 12,
                  metric = "ROC",
                  trControl = trainControl(
                 method = "cv", # k-fold cross validation
                 number = 10, # 10 folds
                 savePredictions = "final",
                 classProbs = TRUE,
                 summaryFunction = defaultSummary
                 )
                  )
rf_fit
```

```
## Random Forest
##
## 297 samples
##  11 predictor
##   3 classes: 'Insomnia', 'None', 'Sleep.Apnea'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 268, 266, 268, 267, 266, 268, ...
## Resampling results across tuning parameters:
##
##   mtry  splitrule   Accuracy   Kappa
##    2    gini        0.9090434  0.8413850
##    2    extratrees  0.9156025  0.8528083
##    3    gini        0.9057100  0.8356049
##    3    extratrees  0.9123767  0.8471099
##    4    gini        0.9022618  0.8296133
##    4    extratrees  0.9057100  0.8356049
##    5    gini        0.9057100  0.8356049
##    5    extratrees  0.9089359  0.8413033
##    6    gini        0.9054876  0.8353118
##    6    extratrees  0.9057100  0.8356049
##    7    gini        0.9022618  0.8296133
##    7    extratrees  0.9022618  0.8296133
##    8    gini        0.9022618  0.8296133
##    8    extratrees  0.9057100  0.8356049
##    9    gini        0.9054876  0.8353118
##    9    extratrees  0.9022618  0.8296133
##   10    gini        0.9054876  0.8353118
##   10    extratrees  0.9055951  0.8353934
##   11    gini        0.9022618  0.8296133
##   11    extratrees  0.9054876  0.8353118
##   12    gini        0.9055951  0.8353934
##   12    extratrees  0.9022618  0.8296133
##
## Tuning parameter 'min.node.size' was held constant at a value
of 1
## Accuracy was used to select the optimal model using the
largest value.
## The final values used for the model were mtry = 2, splitrule =
```

```
extratrees
##   and min.node.size = 1.
```

```
plot(rf_fit)
```

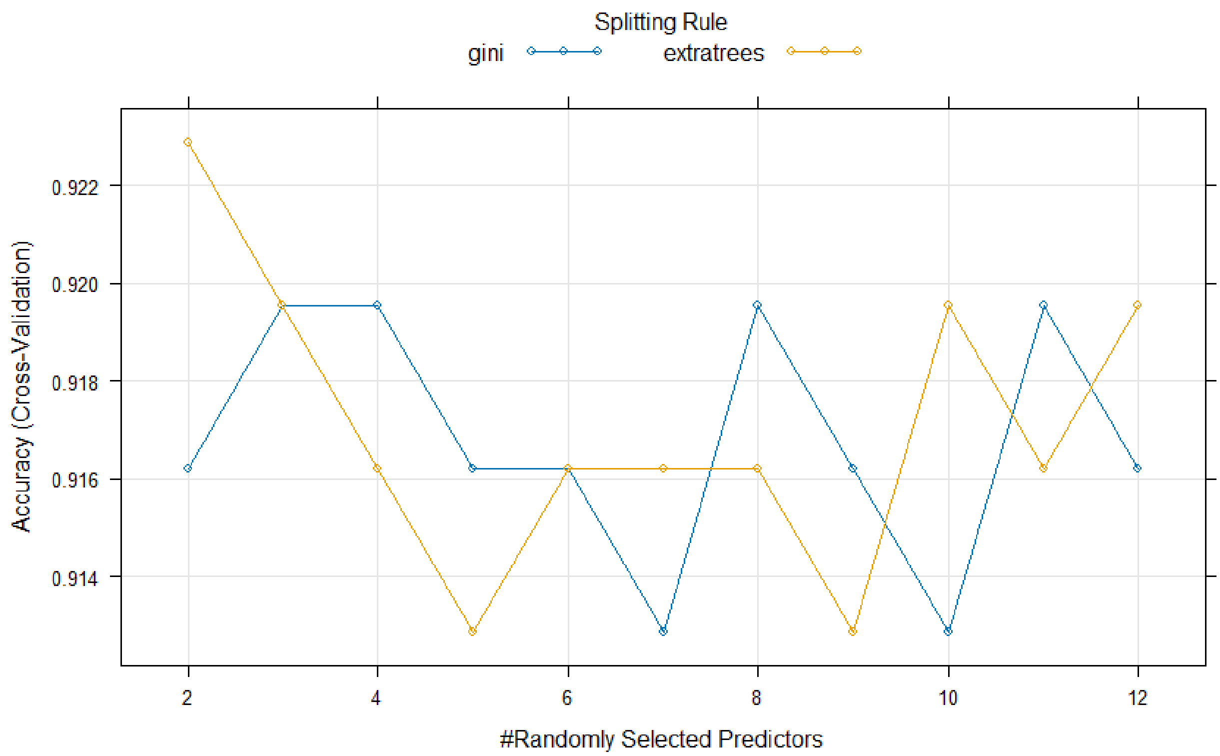

Figure 27. Tune Random Forest Parameters

The mtry = 2, splitrule = extratrees and min.node.size = 1 indicates the accuracy achieves the greatest level.

Step 5: Build the model with the best mtry

```
set.seed(111)
rf_model_2 <- randomForest(Sleep.Disorder~., data = train_data, mtry = rf_fit$bestTune$mtry, splitrule = rf_fit$bestTune$splitrule, importance = TRUE, ntree = 1000)
rf_model_2
```

```
##
## Call:
##  randomForest(formula = Sleep.Disorder ~ ., data = train_data, mtry = rf_fit$bestTune$mtry, splitrule = rf_fit$bestTune$splitrule,      importance = TRUE, ntree = 1000)
```

```
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 8.42%
## Confusion matrix:
##             Insomnia None Sleep.Apnea class.error
## Insomnia          48    5           6  0.18644068
## None               1  166           5  0.03488372
## Sleep.Apnea        5    3          58  0.12121212
```

Step 6: Save model

```
save(rf_model_2, file = "RFModel.rda")
```

6.5 Validation

6.5.1. Test validation of the Decision Tree

```
predict_decision_tree  <-  predict(decision_tree,  newdata  =
test_data,                                        type="class")
confusionMatrix(test_data$Sleep.Disorder, predict_decision_tree)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Insomnia None Sleep.Apnea
##   Insomnia          16    2           0
##   None               4   43           0
##   Sleep.Apnea        1    2           9
##
## Overall Statistics
##
##                Accuracy : 0.8831
##                  95% CI : (0.7897, 0.9451)
##     No Information Rate : 0.6104
##     P-Value [Acc > NIR] : 1.098e-07
##
##                   Kappa : 0.7857
##
```

```
##  Mcnemar's Test P-Value : 0.2998
##
## Statistics by Class:
##
##                      Class: Insomnia Class: None Class:
Sleep.Apnea
## Sensitivity                   0.7619      0.9149
1.0000
## Specificity                   0.9643      0.8667
0.9559
## Pos Pred Value                0.8889      0.9149
0.7500
## Neg Pred Value                0.9153      0.8667
1.0000
## Prevalence                    0.2727      0.6104
0.1169
## Detection Rate                0.2078      0.5584
0.1169
## Detection Prevalence          0.2338      0.6104
0.1558
## Balanced Accuracy             0.8631      0.8908
0.9779
```

6.5.2. Test validation of the Random Forest

```
predict_rf <- predict(rf_model_2, newdata = test_data, type =
"class")
confusionMatrix(test_data$Sleep.Disorder,predict_rf)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Insomnia None Sleep.Apnea
##   Insomnia          16    2           0
##   None               3   44           0
##   Sleep.Apnea        1    2           9
##
## Overall Statistics
##
##                Accuracy : 0.8961
##                  95% CI : (0.8055, 0.9541)
```

```
##     No Information Rate : 0.6234
##     P-Value [Acc > NIR] : 7.181e-08
##
##                   Kappa : 0.8078
##
##  Mcnemar's Test P-Value : 0.3618
##
## Statistics by Class:
##
##                      Class: Insomnia Class: None Class:
Sleep.Apnea
## Sensitivity                   0.8000      0.9167
1.0000
## Specificity                   0.9649      0.8966
0.9559
## Pos Pred Value                0.8889      0.9362
0.7500
## Neg Pred Value                0.9322      0.8667
1.0000
## Prevalence                    0.2597      0.6234
0.1169
## Detection Rate                0.2078      0.5714
0.1169
## Detection Prevalence          0.2338      0.6104
0.1558
## Balanced Accuracy             0.8825      0.9066
0.9779
```

CHAPTER 7: CONCLUSION

Research shows that sleep disorders greatly affect the lives of people of different ages, genders, and health conditions.

For people with sleep apnea syndrome, most women are more likely to have it than men. People with high stress levels and high blood pressure are also a factor leading to sleep apnea syndrome. And another factor that causes sleep apnea is BMI. According to research, obese subjects will suffer from sleep apnea syndrome.

For people with insomnia, the average is higher in women than in men. Lack of physical activity also has an impact on insomnia. Compared to people with a high number of steps of physical activity, people with a lower number of steps of physical activity have a higher risk of insomnia. And people with high stress levels and those in the overweight and obese BIM group had high rates of insomnia. Finally, people with high blood pressure are also more likely to suffer from insomnia.

We built a classification model to classify classes according to sleep disorders (None, insomnia, sleep apnea) of hospital patients. Classification helps doctors know which sleep disorder class the patient belongs to in order to give warnings and appropriate treatment regimens.