Bayesian Network of Sleep Quality

Hyungmin Kim

Supervisor: Jinhee Cho

October 20, 2024

Contents

1. Introduction

1.1 Problem Description

In the modern society today, many people are sleep deprived due to work, studies, or other responsibilities. It's not uncommon for individuals to stay awake late into the night, which negatively affects their sleep quality and overall health. This lack of sleep can lead to short-term issues like fatigue and difficulty concentrating in their daily lives, but it can also cause long-term physical and mental health problems which might drastically reduce their quality of life.

One specific example of this issue can be seen in shift workers [1], who often suffer from irregular sleep patterns due to rotating work hours. These workers frequently experience significant health problems, and their quality of life can drastically decline. Not only for them, but also for everyone in the world, ensuring efficient and high-quality sleep is crucial in maintaining their well-being.

Since there are lots of different factors that affect sleep quality and given the complexity of these interactions, it is necessary to employ advanced analytical methods to gain a deeper understanding of how these variables influence sleep quality. In this report, I will use Bayesian Networks to model and assess how each element influences sleep quality. The primary objective of this project is to conduct a comprehensive analysis of the factors influencing sleep quality and develop practical, scientifically grounded methods for improving sleep effectiveness and overall sleep quality.

1.2 System Objective

The objective of this system is to develop a predictive model using a Bayesian Network to analyze the impact of the factors on sleep quality. This will be achieved by leveraging data and implementing the model in R. For additional information, I plan to fine-tune the Bayesian network by referring to related academic papers. Through this model, the system will mathematically and probabilistically evaluate the influence and provide numerical insights. The ultimate goal of the system is to infer the sleep quality of individuals based on their current input conditions. For data, I utilized two Kaggle datasets [2] [3], extracting meaningful information from each and combined them to create a unified model. The detailed information about the specific datasets is provided in *Table 1.1* below.

#	**Variable Name**	**Variable Type**	**Description**
1	User.ID	Discrete	Unique identifier of the user
2	Age	Continuous	Age of the user
3	Gender	Discrete	Gender of the user, "m", "f"

#	Variable Name	Variable Type	Description
4	Sleep.Quality	Continuous	A score representing the quality of sleep, "1", "2", ..., "10"
5	Bedtime	Discrete	The time the user goes to bed, "HH:MM"
6	Wake.up.Time	Discrete	The time the user wakes up, "HH:MM"
7	Daily.Steps	Continuous	Number of steps taken per day
8	Calories.Burned	Continuous	Total calories burned during the day
9	Physical.Activity.Level	Discrete	User's level of physical activity, "high", "low", "medium"
10	Dietary.Habits	Discrete	The user's dietary habits, "healthy", "medium", "unhealthy"
11	Sleep.Disorders	Discrete	Whether the user has any diagnosed sleep disorders, "no", "yes"
12	Medication.Usage	Discrete	Whether the user takes medications, "no", "yes"

Table 1.1: Variable names, types, and descriptions

1.3 Preprocessing

Before creating the Bayesian Network, I examined the datasets and conducted the preprocessing. I checked for any missing values and processed each attribute using appropriate functions to handle discrete and continuous data.

1.3.1 Handling Missing value

```r
na_counts <- colSums(is.na(data_1))
na_counts[na_counts > 0]

# named numeric(0)
```

First I checked for any missing values across all attributes. I confirmed that there are no missing values for any of the attributes.

1.3.2 Feature Encoding

To ensure the data was in the correct format, I utilized the *as.factor()* and *as.numeric()* functions. Specifically, I applied *as.factor()* to convert categorical variables into factors, enabling

more efficient analysis and modeling. Simultaneously, I used *as.numeric()* to convert any continuous variables into numeric format, ensuring they were appropriately represented for subsequent statistical analysis and modeling processes.

2. Constructing the Bayesian Network Model

2.1 Initial BN from bnlearn library

The bnlearn package offers multiple structure learning algorithms as shown in the *Table 2.1* below.

ID	Name	Type
hc	Hill Climbing	Score-based algorithm
tabu	Tabu Search	Score-based algorithm
iamb	Incremental Association Markov Blanket	Constraint-based algorithm
rsmax2	Hybrid HPC	Hybrid
pc.stable	PC	Constraint-based algorithm
gs	Grow-Shrink	Constraint-based algorithm
hpc	Hybrid Parents and Children	Constraint-based algorithm
mmhc	Max-Min Hill Climbing	Hybrid

Table 2.1: Structure Learning Algorithms in bnlearn

I utilized the Hill Climbing algorithm from for this project since it has a flexibility in exploring complex interactions among various factors. Through the score-based optimization feature, I could evaluate and refine the network structure effectively, ensuring that identifying meaningful relationships that influences sleep efficiency.

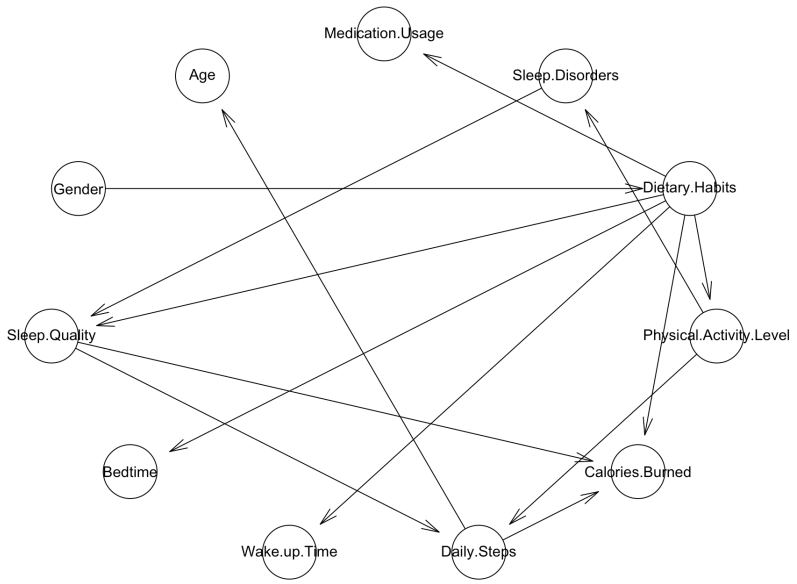

Figure 2.1: (BN 1) Initial Bayesian Network using plot()

This is BN 1, and the resulting model consists of 11 nodes and 14 directed arcs, illustrating the relationships among various factors that influence sleep efficiency. The structure of the Bayesian

network reveals intricate dependencies among these variables. For example, Dietary Habits directly influence Bedtime, Wake Up Time, and Physical Activity Level. Additionally, Physical Activity Level is connected to Sleep Disorders, while both Sleep Quality and Daily Steps are affected by various combinations of Dietary Habits and Sleep Disorders.

Currently, the relationships among the factors in the Bayesian network do not appear convincing or sufficiently robust. This observation leads me to believe that further modifications to the network are necessary to enhance its accuracy and reliability. First, to enhance the interpretability of these relationships, I employed the graphviz.plot() for a more intuitive understanding of the connections among the nodes, making it easier to identify how each factor interrelates as shown in *Figure 2.2*.

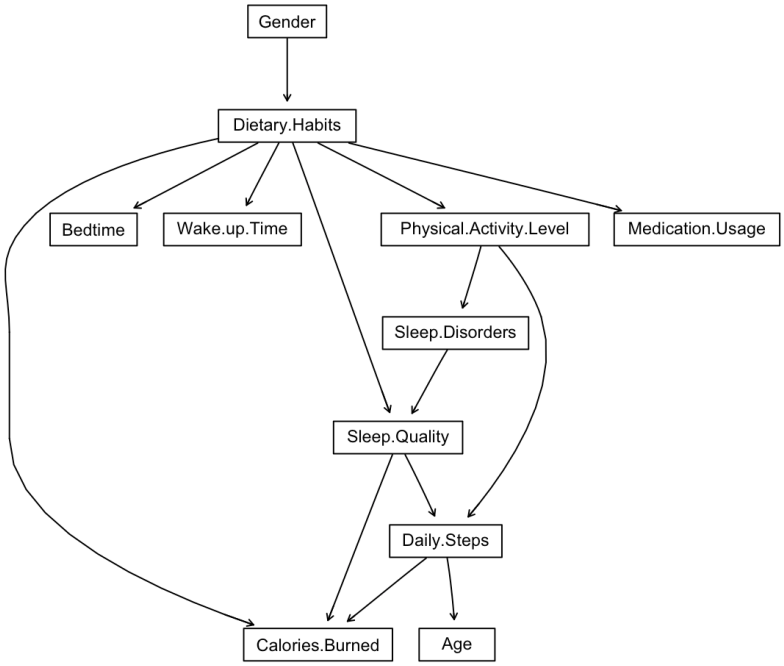

Figure 2.2: (BN 1) Initial Bayesian Network using plot()

Next, I visualized the Mutual Information (MI) using a heatmap to assess the degree of correlation among the factors as shown in below.

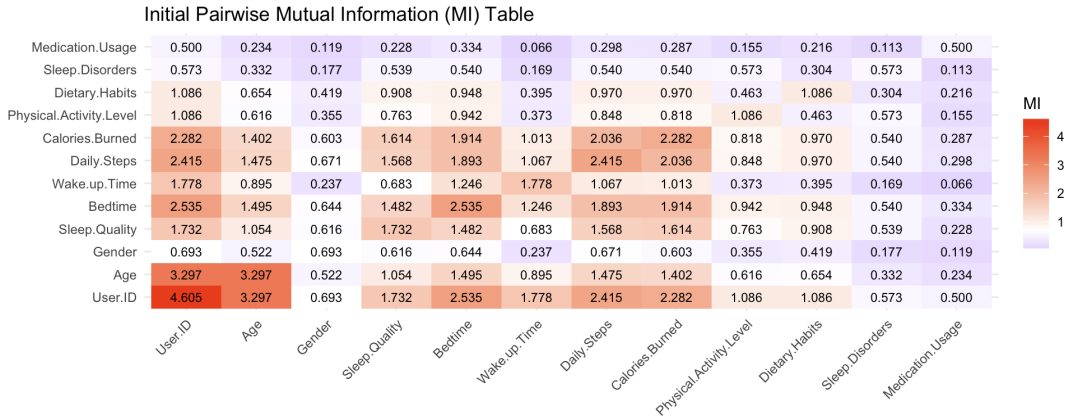

Figure 2.3: MI for BN 1

This MI heatmap illustrates the strength of statistical associations between the factors. The

current data reveals a strong correlation between Daily Steps and Calories Burned, as well as a notable association between Bedtime and Sleep Quality. However, it also contains less meaningful information, such as the very high correlation between User ID and Age, which does not contribute significantly to understanding sleep efficiency. Additionally, it was observed that Medication Usage displayed relatively weak associations with most other variables.

2.2 Adding Variables : Melatonin

Melatonin is a hormone that regulates sleep in humans and affects significantly to the circadian rhythms. While extensive research has been conducted on the impact of melatonin on sleep quality [4] [5], the current dataset lacks this specific data. Therefore, I have decided to utilize another dataset that includes both melatonin levels and sleep quality to construct a model using Bayesian network. Through this model, I aimed to infer the quantity of melatonin level based on the existing data such as sleep duration and age. The variables of the Kaggle dataset was like below, *Table 2.2*.

#	**Variable Name**	**Variable Type**	**Description**
1	Participant. ID	Discrete	Unique identifier for each participant
2	Age	Continuous	Age of the participant
3	Gender	Discrete	Gender of the participant, "m", "f"
4	Chronotype	Discrete	The participant's chronotype, e.g., "morning", "evening"
5	Average.Daily. Social.Media.Use.Time.. minutes.	Continuous	Average time spent on social media daily in minutes
6	Dominant.Social.Media. Platform	Discrete	The most frequently used social media platform
7	Frequency.of.Social.Media. Checking..number.of. times.per.day.	Discrete	Number of times social media is checked per day
8	Pre.Sleep.Social.Media. Use.Duration..minutes.	Continuous	Duration of social media use before sleep in minutes
9	Type.of.Social.Media. Content.Consumed	Discrete	Type of content consumed on social media
10	Sleep.Latency.. minutes.	Continuous	Time taken to fall asleep after going to bed in minutes

#	Variable Name	Variable Type	Description
11	Total.Sleep.Time.. hours.	Continuous	Total sleep duration in hours
12	Sleep.Efficiency....	Continuous	Ratio of total sleep time to time spent in bed
13	Sleep.Quality. Rating	Continuous	A rating representing the quality of sleep, "1" to "10"
14	Wake.After.Sleep. Onset..WASO...minutes.	Continuous	Time spent awake after initially falling asleep in minutes
15	Number.of.Awakenings.. during.sleep.	Discrete	Total number of times the participant wakes up during the night
16	Melatonin.Level.. pg.mL.	Continuous	Melatonin concentration in the blood in picograms per milliliter
17	Cortisol.Level.. pg.mL.	Continuous	Cortisol concentration in the blood in picograms per milliliter
18	Day.of.Week	Discrete	Day of the week during which the data was collected
19	Blue.Light.Exposure. Before.Sleep..minutes.	Continuous	Duration of blue light exposure before sleep in minutes
20	Stress.Level. Rating	Continuous	A rating representing the participant's perceived stress level, "1" to "10"

Table 2.2: Variable names, types, and descriptions

The secretion of melatonin level can be influenced by various factors and one of the significant contributors are age and sleep duration. As people age, the production of melatonin tends to decrease [6], and according to human circadian rhythms, melatonin is secreted most actively between 2 a.m. and 4 a.m [7]. Based on this, I selected Age, Sleep.duration, and Sleep.Quality.Rating as meaningful variables from the second dataset to analyze their correlation with melatonin levels and to extract a BN model for predicting melatonin levels.

To achieve this, it was necessary to standardize the levels of the corresponding factors across both datasets. For instance, in the case of Sleep.duration, the data was discretized into intervals ranging from 3 to 10 hours, with 30-minute increments, by considering the minimum and maximum values of both datasets. Similarly, Age was discretized into intervals of 10 years, ranging from 0 to 100 years, to ensure consistency across the datasets.

Group	Age Range	# of Data
0	0 - 9	0
1	10 - 19	27
2	20 - 29	91
3	30 - 39	89
4	40 - 49	119
5	50 - 59	116
6	60 - 69	58
7	70 - 79	0
8	80 - 89	0
9	90 - 99	0

Table 2.3: Age Groups

Group	Sleep Duration (hours)	# of Data
0	3.0 - 3.5	60
1	3.5 - 4.0	43
2	4.0 - 4.5	68
3	4.5 - 5.0	59
4	5.0 - 5.5	54
5	5.5 - 6.0	50
6	6.0 - 6.5	58
7	6.5 - 7.0	60
8	7.0 - 7.5	31
9	7.5 - 8.0	14
10	8.0 - 8.5	3
11	8.5 - 9.0	0
12	9.5 - 10.0	0

Table 2.4: Sleep Duration Groups (in hours)

In the analysis of melatonin levels, I employed the Hartemink method to create intervals while preserving the interactions between different variables. Although alternative methods such as quantiles and intervals were considered, Hartemink was determined to be the most optimal approach for discretizing the data without losing dependency relationships.

Group	Melatonin Level (pg.mL)	# of Data
0	4.95 - 22.7	263
1	22.7 - 40.4	188
2	40.4 - 58.2	49

Table 2.5: Melatonin Level (pg.mL)

After discretizing all factors, I used the BIC score of the BN 2 model trained on the melatonin data to evaluate before making predictions. The BIC score was calculated using the following R code:

```
bic_score_bn_5 <- score(bn_5, data = data_5_df, type = "bic")
```

The resulting BIC score was: **BIC Score: -1122.31662301879**

A lower BIC score suggests a better model, penalizing excessive parameters to avoid overfitting. In this case, a score of -1122.3166 indicates a strong fit for the data while maintaining parsimony. This score can guide the selection of appropriate model among competing structures, ensuring accuracy and complexity in predicting melatonin levels. The graph of the BN 2 is below *(Figure 2.4.)*

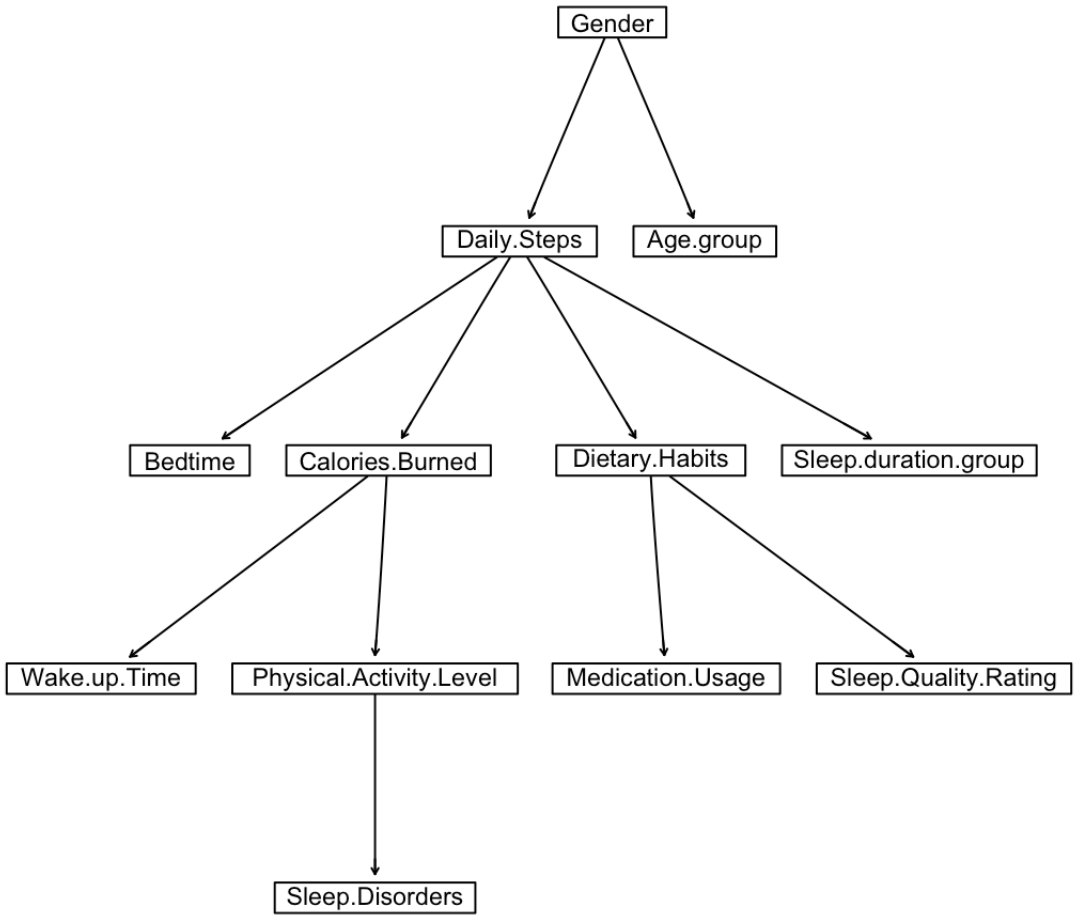

Figure 2.4: (BN 2) BN after discretizing all factors

Using the fitted model, I utilized Likelihood Weighting to infer melatonin levels from the original dataset. While considering other methods such as exact, approximate, and forward inference methods, likelihood weighting was chosen due to its ability to perform complex evidence-based inference quickly. This method can efficiently provide approximate values through sampling, maintaining adequate performance in complex networks, and achieving high accuracy with sufficient samples. Missing values encountered during this process were addressed using kNN for imputation.

```
data_7 <- kNN(data_6)
any(is.na(data_7))
```

The Bayesian network used to infer melatonin levels is presented below at the *Figure 2.5.*. It was confirmed that melatonin levels directly influence the Sleep Quality Rating, and relationships with other factors are also illustrated in the *Figure 2.5.*. There are 12 directed arcs and 13 nodes, with an average Markov blanket and neighborhood size of 1.85, suggesting a well-connected structure. While interpreting BN, I could observe meaningful interpretations like dietary habits are influenced by Daily Steps, which in turn affect other variables like Sleep Quality Rating and Medication Usage.

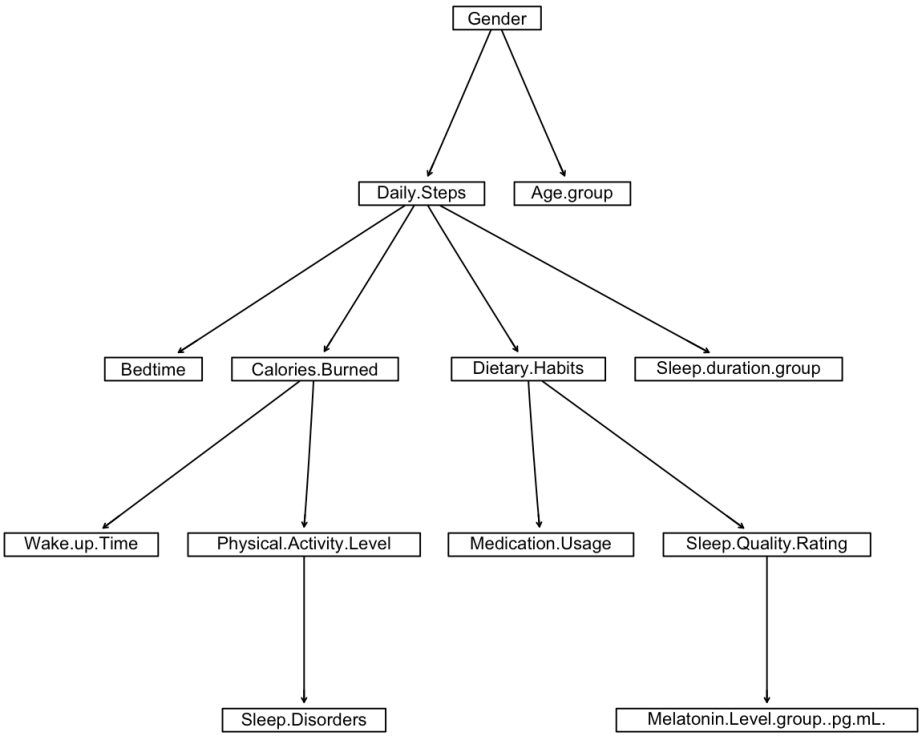

Figure 2.5: (BN 3) BN after discretizing all factors

The data counts for the inferred melatonin levels are summarized in the *Table 2.6.*.

Group	Melatonin Level (pg.mL)	# of Data
0	(-Inf) - 22.7	14
1	22.7 - 40.4	44
2	40.4 - (Inf)	42

Table 2.6: Melatonin Level (pg.mL) for BN 3

2.3 Complexity Reduction

2.3.1 Mutual Information

During the comparison between MI from BN 3 with that from BN 1, I could find significantly more informative insights. For instance, the MI between Bedtime and Wake up Time is 2.535, indicating a strong correlation between these two variables. Similarly, the MI of 2.023 between Sleep Duration Group and Age Group suggests substantial differences in sleep duration across age categories.

Notably, the MI of 0.995 between Sleep Quality Rating and Melatonin Level highlights a strong association, emphasizing the impact of melatonin on sleep quality. Additionally, I observed relationships such as the MI of 1.086 between Physical Activity Level and Dietary Habits, and

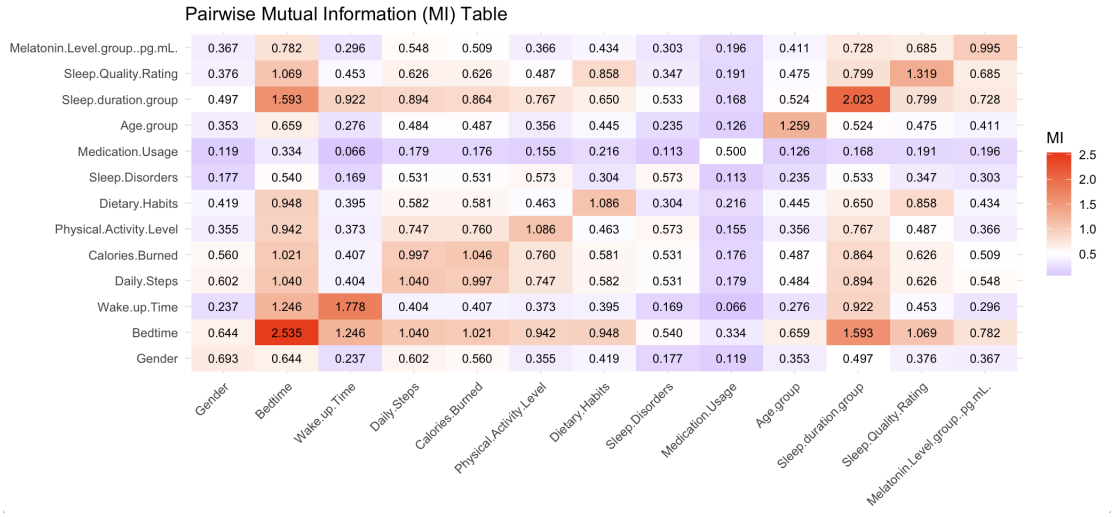

Figure 2.6: MI for BN 3

approximately 1.0 between Calories Burned and Daily Steps. In contrast, Medication Usage showed low correlations with most variables (MI < 0.5), suggesting its relative independence from other sleep-related factors.

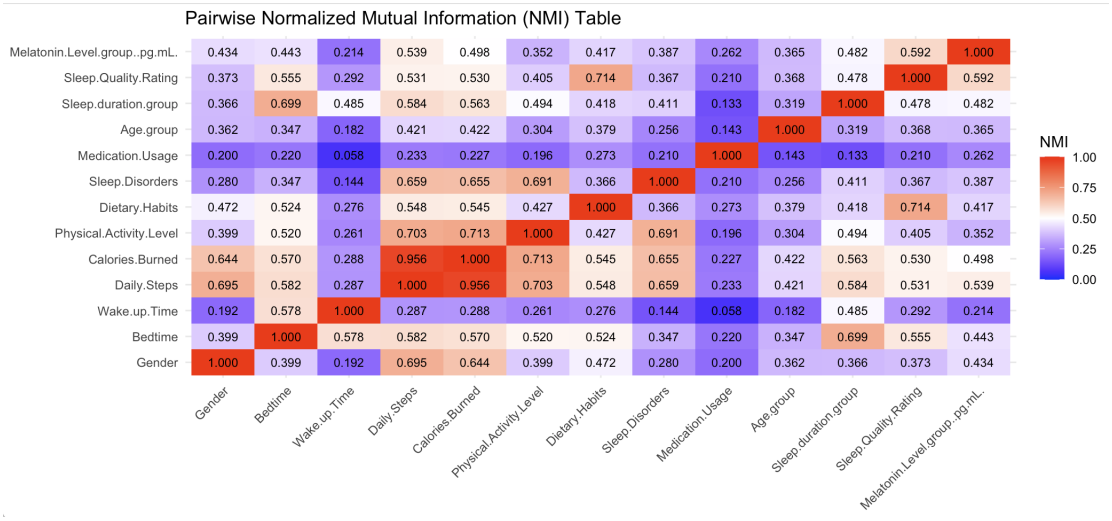

Figure 2.7: MI for BN 3

Then I normalized the mutual information (MI) for each factor and created a Normalized Mutual Information (NMI) table. This analysis revealed that sleep patterns, such as Bedtime, Wake up Time, and Sleep Duration, are strongly interconnected. Additionally, lifestyle factors, including Physical Activity, Dietary Habits, and Calories Burned, show significant correlations. Melatonin Levels are closely related to overall Sleep Quality, while Age demonstrates a substantial relationship with Sleep Duration. These findings highlight the intricate relationships among various factors influencing sleep.

2.3.2 Whitelist and Blacklist

I made further adjustments to the BN 3 by leveraging the whitelist and blacklist features of the bnlearn package. Drawing on expert opinions and academic journals, I evaluated how the values of certain variables impact others factors and revised some prior and conditional probabilities accordingly. The established whitelist and blacklist are detailed in the table below.

From	To
Gender	Sleep Disorders
Dietary Habits	Sleep Quality Rating
Dietary Habits	Sleep Duration Group
Physical Activity Level	Sleep Quality Rating
Wake up Time	Melatonin Level (group..pg.mL.)
Bedtime	Melatonin Level (group..pg.mL.)
Physical Activity Level	Melatonin Level (group..pg.mL.)
Age Group	Melatonin Level (group..pg.mL.)

Table 2.7: Whitelist

From	To
Age Group	Gender

Table 2.8: Blacklist

Gender and Sleep Disorders

Research has shown that intensity of sleep disorders can vary between males and females due to biological, hormonal, and psychological factors. Specifically, women are more prone to insomnia, restless leg syndrome, and other sleep disturbances, partly due to hormonal cycles and life events like menopause [8].

Dietary Habits and Sleep Quality Rating

Dietary habits play a crucial role in determining sleep quality. Studies indicate that certain foods and eating patterns such as caffeine or vitamin intake can significantly impact the sleep quality [9].

Wake Up Time, Bedtime and Melatonin Level

Melatonin secretion naturally increases in response to darkness and peaks during the night. Going to bed at irregular or late hours can disrupt this rhythm, leading to reduced melatonin levels, poorer sleep quality, and difficulty falling asleep [10].

Physical Activity Level and Melatonin Level

Physical activity positively influences melatonin production, enhancing sleep quality. Regular exercise increases melatonin secretion, helping individuals fall asleep faster and sleep more soundly by regulating the sleep-wake cycle [11].

Age Group and Melatonin Level

Melatonin levels naturally decline with age, affecting sleep quality and circadian rhythm regulation, particularly in older adults. This reduction may contribute to lack of sleep quality and weakened immune function [12][13].

Age and Gender

The decision to include gender in the blacklist was made based on the observation that the gender distribution in the dataset is precisely 50/50. This indicates that age does not influence gender and allows for a more accurate representation of the data without unnecessary complexity. Consequently, this adjustment ensures that gender remains independent of age-related variables in the Bayesian network.

By Adjusting whitelist and blacklist, I could get BN 4.

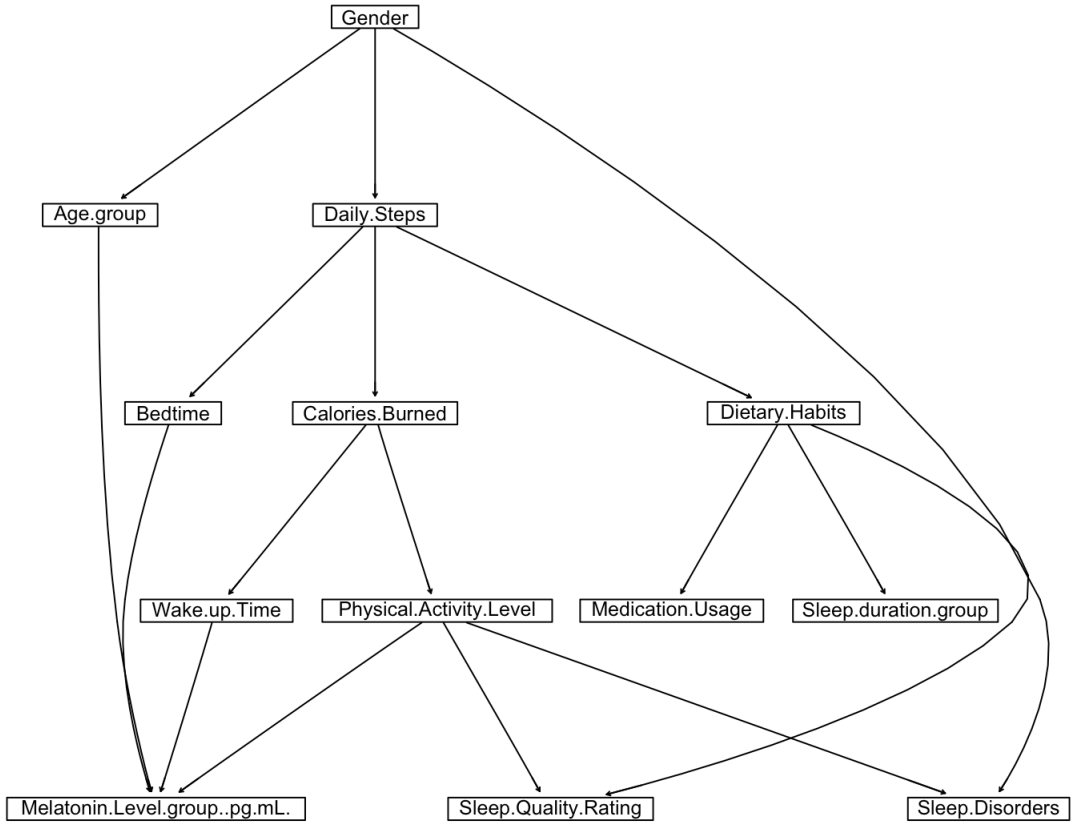

Figure 2.8: (BN 4) BN after whitelist and blacklist

2.3.3 Comparison of Independent Parameters

For the convenience of calculating the number of independent parameters, the initials of each factor are represented as follows: "Age" is denoted as A, "Gender" as G, "Sleep Quality" as Q, "Bedtime" as B, "Wake Up Time" as W, "Daily Steps" as S, "Calories Burned" as C, "Physical Activity Level" as P, "Dietary Habits" as H, "Sleep Disorders" as D, and "Medication Usage" as U.

Before constructing the Bayesian model, the independent parameters calculated from the initial dataset (excluding UserID) yield:

$29(A)\times 2(G)\times 10(Q)\times 14(B)\times 7(W)\times 15(S)\times 11(C)\times 3(P)\times 3(H)\times 2(D)\times 2(U)-1 = \mathbf{337{,}629{,}599}$

independent parameters.

After structuring the initial Bayesian Network BN 1, the number of states for each node remains the same, while the parent counts for each node are as follows:

Node	States	Parents' States	Parameters
(G) Gender	2	1	1
(H) Dietary Habits	3	2	4
(B) Bedtime	14	3	39
(W) Wake Up Time	7	3	18
(P) Physical Activity Level	3	3	6
(U) Medication Usage	2	3	3
(D) Sleep Disorders	2	3	3
(Q) Sleep Quality	10	2	18
(S) Daily Steps	15	10	140
(A) Age	29	15	420
(C) Calories Burned	11	450	4500

Table 2.9: Node States, Parents, and Parameters for BN 1

Thus, calculating the independent parameters using the formula:

$$\textbf{Independent Parameters} = \sum(\textbf{Number of States of Node}-1)\cdot(\textbf{Number of Parents' States})$$

we obtain:

$$1+4+39+18+6+3+3+18+140+420+4500 = \mathbf{5{,}152}$$

for BN 1.

Similarly, BN 2, BN 3 and BN 4 can be calculated.

Node	States	Parents' States	Parameters
(G) Gender	2	1	1
(H) Dietary Habits	3	3	6
(B) Bedtime	14	3	39
(W) Wake Up Time	7	3	18
(P) Physical Activity Level	3	3	6
(U) Medication Usage	2	3	3
(D) Sleep Disorders	2	3	3
(Q) Sleep Quality	4	3	9
(S) Daily Steps	3	2	4
(A) Age	4	2	6
(C) Calories Burned	3	3	6

Table 2.10: Node States, Parents, and Parameters for BN 2

BN 2:

$$1+6+39+18+6+3+3+9+4+6+6 = \mathbf{101}$$

Node	States	Parents' States	Parameters
(G) Gender	2	1	1
(H) Dietary Habits	3	3	6
(B) Bedtime	14	3	39
(W) Wake Up Time	7	3	18
(P) Physical Activity Level	3	3	6
(U) Medication Usage	2	3	3
(D) Sleep Disorders	2	3	3
(Q) Sleep Quality	4	3	9
(S) Daily Steps	3	2	4
(A) Age	4	2	6
(C) Calories Burned	3	2	6
(M) Melatonin Level	3	4	8

Table 2.11: Node States, Parents, and Parameters for BN 3

BN 3:

$$1 + 6 + 39 + 18 + 6 + 3 + 3 + 9 + 4 + 6 + 6 + 8 = \mathbf{109}$$

Node	States	Parents' States	Parameters
(G) Gender	2	1	1
(H) Dietary Habits	3	3	6
(B) Bedtime	14	3	39
(W) Wake Up Time	7	3	18
(P) Physical Activity Level	3	3	6
(U) Medication Usage	2	3	3
(D) Sleep Disorders	2	6	6
(Q) Sleep Quality	4	9	27
(S) Daily Steps	3	2	4
(A) Age	4	2	6
(C) Calories Burned	3	3	6
(M) Melatonin Level	3	1176	2352

Table 2.12: Node States, Parents, and Parameters for BN 4

BN 4:

$$1 + 6 + 39 + 18 + 6 + 3 + 6 + 27 + 4 + 6 + 6 + 2352 = \mathbf{2{,}474}$$

In conclusion, by reducing the initial number of independent parameters in the Bayesian model from 5,152 to 2,474, the number of parameters was decreased by approximately 52.0%. Compared to BN 3, the number of independent parameters in BN 4 increased from 109 to 2,474. This significant rise in the parameter count can be attributed primarily to the specification that melatonin levels are influenced by Bedtime and Wake Up Time through the use of a whitelist.

This increase leads to greater model complexity, allowing for more flexible modeling and more precise predictions. Additionally, it can be observed that the accuracy of inferences has improved. In cases of high-dimensional data, the model is likely to demonstrate even better performance.

2.4 Inference Analysis

As mentioned at the beginning of the report, the final goal of this project is to infer the probability of a (Q) Sleep Quality based on the various sets of evidence using the constructed latest Bayesian Networks and computed prior probabilities. In this project, I will use the cpquery function from the bnlearn package in R in this order:

```
cpquery(fitted, event, evidence, cluster, method = "ls", ...,
  debug = FALSE)
# cpquery estimates the conditional probability of event given evidence
    using the method specified in the method argument.
```

1. **event**: Define the event (the condition that want to predict).
2. **evidence**: Specify the evidence (the conditions that already known).
3. **query**: Call the cpquery function to compute the probability.

For example, we can infer the probability of whether "(Q) Sleep Quality" is equal to 4, who has age between 20 and 29, is a male, dietary habits are not healthy, slept at 1:15 am and who has sleep disorders by setting like below.

```
bn_8_model <- bn.fit(bn_8, data = data_8)
predicted <- cpquery(bn_8_model, event = (Sleep.Quality.Rating=="4"),
    evidence =(Age.group == "20-29" &

               Gender == "m" &

               Dietary.Habits == "unhealthy" &

               Bedtime == "01:15" &

               Sleep.Disorders == "yes"

               ))
predicted
# predicted = 0.03225806
```

The inferred result said there is 3.23% of probability that the person might have Sleep Quality Rating in category '4'. You can use other categories and evidences to infer variables and other factors.

2.5 Uncertainty Factors

In this project, the uncertainty factors considered in the constructed BN include demographic data, lifestyle choices and health conditions which can not be measured precisely. In addition, the goal of the project, Sleep quality ratings are subjective and not easily measurable, making it challenging to establish clear relationships with other factors. To address and minimize these uncertainties, I leveraged multiple dataset, researched lots of academic journals, and compared and employed various methods. Although I could not fully utilize all the data from the multiple

datasets (such as smartphone usage before sleep time and light brightness), I believe that I significantly reduced uncertainty by inferring melatonin data, which has the greatest impact on sleep quality. Also, I referred numerous studies that explain the relationships among the various factors. Finally, I selected the optimal algorithm to reduce uncertainty by considering various methods for discretizing continuous variable data and the algorithms used to construct the BN structure.

3. Results

3.1 Evaluation of Inference Accuracy and Algorithmic Complexity

3.1.1 Effectiveness

I split the train and test datasets in an 8:2 ratio for model training. Due to the limited number of data points (100 total), I applied data augmentation to add 300 more data points, resulting in a total of 380 training samples.

For the training dataset, the accuracy, recall, precision, and F-score were calculated for the five classes of **Sleep Quality Rating** ("1", "2", "3", "4", "5"), with the following results:

- **Accuracy**: 0.8181818
- **Recall**: NA 0 1 1 1
- **Precision**: NA NA 0.3333333 1 1
- **F-score**: NA NA 0.5 1 1

The accuracy shows a fairly strong performance, with approximately 81.82% of the overall predictions being correct. In terms of recall, the value for the first class is NA, which is due to the absence of any samples in the dataset where the Sleep Quality Rating is 1. This highlights a potential issue with data insufficiency for this class, leading to low performance of the model. The second class also has no recall value (0), while the recall for the third, fourth, and fifth classes is 1, indicating that the model was able to correctly identify all positive samples in these classes. For precision, the first two classes (1 and 2) have NA values. The F-score similarly reflects the model's uneven performance across different classes. While it is also not computable for the first two classes, it is 0.5 for the third class and 1 for the fourth and fifth classes.

In summary, while the model performs well in certain classes, its effectiveness is inconsistent across all classes. Particularly, the low recall and precision for some classes indicate a need for more balanced training data, especially for the first and second classes.

3.1.2 Efficiency

In general, the time complexity in Bayesian networks is primarily determined by two factors:

1. Network structure learning

2. Inference algorithms

For network structure learning, the Hill Climbing algorithm was employed. The time complexity of this approach is approximately $O(n^2)$ for neighbor search, where n represents the number of variables in the network. Although the time complexity is also influenced by the number of variables and the size of the parent node sets, we increased efficiency by applying whitelist and blacklist constraints.

The inference algorithm used in this project is Likelihood Weighting, a type of Monte Carlo method. While this method can be time-consuming due to the large number of samples required, it is an efficient approach that focuses on approximations rather than exact probability computations.

3.2 Sensitivity Analysis

During this project, I constructed multiple BN structures to optimize the model's performance. To capture the relationships between the variables, I also have applied various preprocessing techniques such as the discretization of continuous variables and grouping of key features. Variables such as Age, Sleep.duration, Daily.Steps, and Calories.Burned were transformed from continuous to categorical to better fit the model. Additionally, a new dataset, including the Melatonin.Level variable, was integrated into the BN construction. This process allowed the model to learn significant relationships between variables. By constructing multiple network structures, we built a strong foundation for conducting sensitivity analysis, ensuring that interactions between variables are captured accurately.

Through the iterative construction of various BN models, I have made a strong foundation to evaluate the impact of different variables on the target variable. The sensitivity analysis can help quantify interactions between variables and identify the most influential variables, thereby improving the predictive accuracy of future models.

3.3 Novelty and Challenge

The goal of this project was focused on the subjective and complex topic of sleep quality, which poses significant challenges in quantification and analysis. The purpose of choosing topic was to aim uncover unexpected correlations between variables through the implementation of a Bayesian Network. This innovative approach allowed for a deeper exploration of the intricate relationships within the data, highlighting aspects of sleep quality that are often overlooked.

Throughout the project, I conducted various extensive hypothesis formulation and validation, taking thorough research to support each proposed idea. The iterative nature of this process required not only critical thinking but also adaptability as new data insights emerged. Implementing these hypotheses necessitated the integration of various datasets and techniques, reflecting the project's multifaceted nature.

One of the notable innovative aspects was addressing the limitations posed by insufficient data. I employed different methods to augment and combine multiple datasets, enhancing the overall robustness of the analysis. Although the absence of additional data like adenosine [1] and light condition [14][15] lower the accuracy of the model and prevented a comprehensive validation of the model, the project exemplified a resourceful application of available assets to derive optimal conclusions under constrained conditions.

The findings demonstrated that all factors within the utilized datasets are indeed associated with sleep quality, resulting in meaningful insights. This project serves as a significant contribution to understanding the complexities surrounding sleep quality, combining rigorous analysis with creative problem-solving.

4. References

[1] Reichert, Carolin Franziska, Micheline Maire, Christina Schmidt, and Christian Cajochen. "Sleep-Wake Regulation and Its Impact on Working Memory Performance: The Role of Adenosine." Biology 5, no. 1 (March 2016): 11. https://doi.org/10.3390/biology5010011

[2] Han Aksoy, "Health and sleep statistics", https://www.kaggle.com/datasets/hanaksoy/health-and-sleep-statistics

[3] Global Media Data, "SocialMediaUsage SleepData SG", https://www.kaggle.com/datasets/globalmediadata/socialmediausage-sleepdata-sg

[4] Atul Khullar, M. D. "The Role of Melatonin in the Circadian Rhythm Sleep-Wake Cycle," Psychiatric Times Vol 29 No 7, 29 (July 9, 2012). https://www.psychiatrictimes.com/view/role-melatonin-circadian-rhythm-sleep-wake-cycle

[5] "New Perspectives on the Role of Melatonin in Human Sleep, Circadian Rhythms and Their Regulation - PMC." Accessed October 21, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC6057895/

[6] Burgess, Helen J., and Louis F. Fogg. "Individual Differences in the Amount and Timing of Salivary Melatonin Secretion." PLoS ONE 3, no. 8 (August 26, 2008): e3055. https://doi.org/10.1371/journal.pone.0003055

[7] Kennaway, David J. "The Dim Light Melatonin Onset across Ages, Methodologies, and Sex and Its Relationship with Morningness/Eveningness." Sleep 46, no. 5 (May 1, 2023): zsad033. https://doi.org/10.1093/sleep/zsad033

[8] Andersen, Monica L., Helena Hachul, Isabela Antunes Ishikura, and Sergio Tufik. "Sleep in Women: A Narrative Review of Hormonal Influences, Sex Differences and Health Implications." Frontiers in Sleep 2 (December 19, 2023). https://doi.org/10.3389/frsle.2023.1271827

[9] Sejbuk, Monika, Iwona Mirończuk-Chodakowska, and Anna Maria Witkowska. "Sleep Quality: A Narrative Review on Nutrition, Stimulants, and Physical Activity as Important Factors." Nutrients 14, no. 9 (January 2022): 1912. https://doi.org/10.3390/nu14091912

[10] Zisapel, Nava. "New Perspectives on the Role of Melatonin in Human Sleep, Circadian Rhythms and Their Regulation." British Journal of Pharmacology 175, no. 16 (January 15, 2018): 3190. https://doi.org/10.1111/bph.14116

[11] Alnawwar, Majd A., Meiral I. Alraddadi, Rafaa A. Algethmi, Gufran A. Salem, Mohammed A. Salem, Abeer A. Alharbi, Majd A. Alnawwar, et al. "The Effect of Physical Activity on Sleep Quality and Sleep Disorder: A Systematic Review." Cureus 15, no. 8 (August 16, 2023). https://doi.org/10.7759/cureus.43595

[12] Mayo Clinic. "Melatonin." Accessed October 22, 2024. https://www.mayoclinic.org/drugs-supplements-melatonin/art-20363071

[13] Karasek, M. "Melatonin, Human Aging, and Age-Related Diseases." Experimental Gerontology 39, no. 11–12 (2004): 1723–29. https://doi.org/10.1016/j.exger.2004.04.012

[14] Burgess, Helen J., and Charmane I. Eastman. "Early versus Late Bedtimes Phase Shift the Human Dim Light Melatonin Rhythm despite a Fixed Morning Lights on Time." Neuroscience Letters 356, no. 2 (February 12, 2004): 115. https://doi.org/10.1016/j.neulet.2003.11.032

[15] Brennan, R., J. E. Jan, and C. J. Lyons. “Light, Dark, and Melatonin: Emerging Evidence for the Importance of Melatonin in Ocular Physiology.” Eye 21, no. 7 (July 2007): 901–8. https://doi.org/10.1038/sj.eye.6702597.