JOURNAL OF DATA SCIENCE
eISSN:2805-5160
Vol.2022:02

Uncovering Relationship between Sleep Disorder and Lifestyle using Predictive Analytics

Sham Fang Ying[1], Harprith Kaur[1*], Deshinta Arrova Dewi[1], Chong Fong Kim[1]
Faculty of Information Technology, INTI International University, Malaysia

Email: i19017371@student.newinti.edu.my, harprith.randhawa@newinti.edu.my, deshinta.ad@newinti.edu.my

Sleep disorder refers to the conditions that affect the ability of someone to sleep well regularly whether they are caused by health problems or other outside influences. Occasionally most people experience a sleeping problem due to various reasons. However, when this issue keeps occurring and interferes with daily life, this may indicate a sleeping disorder. In some cases, a sleep disorder may be a symptom of another medical or mental health condition and eventually gone once treatment is obtained for the underlying cause. The treatment normally involves a combination of medical treatments and lifestyle changes. Previous research reported that someone's lifestyle may affect the sleep length and its quality. For example, food choice affects sleep quality and caffeine consumption affects sleep length. This paper aims to uncover the relationship between sleep disorder and lifestyle by performing data investigation using predictive analytics. This study employs Cross Industry Standard Process for Data Mining (CRISP-DM) as methodology. Starting with collection of raw datasets, which were acquired from SleepFoundation.org, one of the leading sources of evidence-based pertaining sleep health information. From there, 1000 data records with 77 attributes are selected and categorized into five class labels i.e. Personal, Diet, Technology, Disease, and Environment. The 77 attributes including depression, anxiety disorder, felt sad, overall health, etc. are then measured using Cramer's value and visualize using Mosaic plots. The Correlation Coefficient and P-value methods are employed to define the relationship among those attributes with a sleep disorder. As for the predictive analytics, we exploit three data mining methods i.e. Support Vector Machine (SVM), Conditional Inference Tree (CTree) and Recursive Partitioning (Rpart). Results show that SVM lead the accuracy level up to 80.288% outperformed Rpart (71.428%) and Ctree (66.499%).

Keywords: ***Sleep Disorder, Data Mining, Predictive Analytics***, ***Machine Learning***

1. Introduction

Sleep is the gold chain that binds health and bodies together. Sleep is a crucial part of daily routine, it takes up around one-third of human daily life. Sleep quality is just as important as food and water for life. Humans cannot develop or maintain the neural connections in their brains that enable them to learn and generate new memories if they do not get enough sleep. It leads to difficulties concentrating and respond fast. Sleep is necessary for a variety of brain processes, including the communication between nerve cells (neurons). In reality, when we sleep, the brain and body remain incredibly active. Sleep serves as a housekeeper that removes toxins from a brain that accumulate while we are awake.

Submission: 17 October 2021
Acceptance: 11 March 2022

Sleep disorders are kinds of conditions that make people unable to have their regular sleep habits. There are about 80 distinct types of sleep disorders. Some of the most common kinds are:

- Insomnia: An inability to fall and remain asleep. It is the most prevalent sleep problem.
- Sleep apnoea: A respiratory disorder in which someone stops breathing for at least 10 seconds while sleeping. According to Donovan and Kapur (2016) research, 0.9% of 40 years old males have this condition.
- Restless leg syndrome (RLS): A tingling or prickly feeling in the legs, patients will have a strong desire to move them. 5-10% of adults and 2-4% of children have suffered from this condition (MedlinePlus, 2020).
- Hypersomnia: A condition, which a person is unable to stay awake during the day.
- Circadian rhythm disorders: Issues on the sleep-wake cycle. It makes the patients unable to sleep and wake at a regular time.
- Parasomnia: An act of walking, talking, or eating while falling asleep, sleeping, or waking up. 66% of people have experience talking while they are asleep.

The previous research of Campsen and Buboltz (2017) stated many reasons for lifestyle affect sleep's length and quality. The example of their findings showed the food choice was related to one's sleep quality, caffeine consumption was also related to sleep length. The number of hours worked at night decreased the number of hours slept each night. Altogether, it concludes that many lifestyles and actions in the day that affect a person's sleep during the night. Hence, the study carried out in this paper focuses on how lifestyle and other issues affect sleep quality. We explore the relationship between people's lifestyles with the associated sleep problem. People's inadvertent lifestyles have a great influence on sleep. However, not all people understand which lifestyle habits can indirectly lead to sleep problems. People may know that some lifestyle habits can affect sleep, such as drinking coffee or other lifestyle habits that cause sleep disorders. It is then understood when the National Centre for Biotechnology Information (2016) reported by 10% to 30% of adults struggle with chronic insomnia. Besides, around 30 % to 48 % of older adults suffer from insomnia (Patel, Steinberg & Patel, 2018).

Sleep problems can be observed through daily behavior because they lead to some concerns such as unable able to concentrate on tasks completion, sluggish and weak. Those with severe sleep problems may lead to accidents in the workplace or while driving. According to the Centres for Disease Control and Prevention (2020), drowsy driving is responsible for more than 6,000 fatal car crashes every year in the United States. If people do not have enough sleep to drive on the road, it is very dangerous. This will not only endanger their safety even endanger the safety of others on the road. In some cases, sleep problems may cause minor memory loss.

According to Cleveland Clinic medical professionals (2020), people can observe the sleep problems through some of the questions like whether they fell asleep while driving, difficulty to pay attention, performance dropped, difficulty to memory, slowed responses to the environment, difficulty to control their emotions, etc. Although the causes of sleep problems may vary, the result of all sleep disorders is that the body's natural sleep and daytime wakefulness cycles are disrupted. The factors include the physical, medical, psychiatric, environmental, genetics, medications factors, etc. Sleep disorders may not be fatal, but they frequently and severely affect people's quality of life. They can even disrupt people's thinking, weight, academic or work performance, mental health as well as general physical health.

Submission: 17 October 2021
Acceptance: 11 March 2022

We formulae the three research questions (RQ) and three research objectives (RO) as follows:
RQ1 – What is the relationship between lifestyle and sleep problems?
RQ2 – What is the significant lifestyle that will affect sleep problems?
RQ3 – How can people improve their sleep quality by understanding their lifestyle?

RO1 – To identify the correlation between sleep disorders and different lifestyles.
RO2 - To construct a predictive model with the use of different data mining algorithms.
RO3 –To select the most accurate model, which identifies the probability of getting sleep disorders.

2. Methods

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a cycle of methodology that comprises six different phases. The reason CRISP-DM methodology is used in this study is that it is an open standard (IBM, 2020), widely used on the markets, and allows the creation of an adaptive data mining model. CRISP-DM is a flexible model and easy to customize. Figure 1 below depicts the cycle of CRISP-DM.

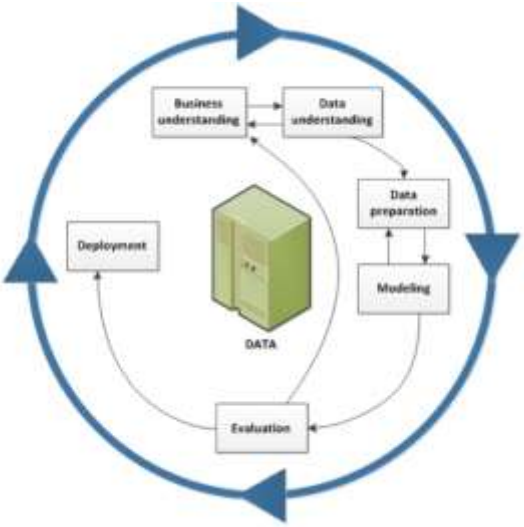

Figure 1. The CRISP-DM life cycle (IBM, 2020)

The first phase is business understanding where we define the problem and objective that we wish to solve in this study. After that, we perform the data-understanding phase where we handle data that aids the solution of the business objective. We collected the initial data and validated whether the data was usable for our analysis concerning the business-understanding phase. In the data preparation stage, we do several activities such as data gathering, data selection, data cleaning, data construction, data integration, etc. This stage is vital for the modelling phase and predictive analytics. Subsequently, in the data preparation phase we reformat the data into another type of value, for example, we change the hours from string values into numeric values, such as “11:00 am” to “11:00” to ease processes in the following phase. Following this is the modelling phase where the selection of modelling techniques was performed, test designs were generated, built, and assessed with the help of machine learning algorithms. Afterward is the evaluation phase whereby all models are validated and checked against the initial requirements in this study. At this stage, the model’s performance in terms of accuracy, validation, etc. is observed. The last phase in CRISP-DM is

Submission: 17 October 2021
Acceptance: 11 March 2022

deployment whereby the actual result of model implementation was successfully conducted and presented to the target audiences.

3. Result and Discussion

In this section, we present the results and discussion of several steps that we performed during the CRISP-DM phases i.e. data cleaning, data reduction, and data mining. Data cleaning is one of the crucial stages in the data analysis process which before that, we had selected 1,000 records with 77 attributes that were categorized into five categories i.e. Personal, Diet, Technology, Disease, and Environment.

In the data cleaning stage, the raw data will be transformed into a different format that can be understood by computers and machine learning. Using box plot visualization we were able to detect outliers of weight and age in the dataset. The box plot is constructed according to the five summaries that consist of the most extreme values in the dataset (also known as the maximum and minimum values), the lower quartile (1st quartile), and upper quartile (3rd quartile), and the median. The five summaries are useful in descriptive analysis during the preliminary investigation of a large dataset. Figure 2 illustrates how the boxplot was able to remove the outliers for weight data.

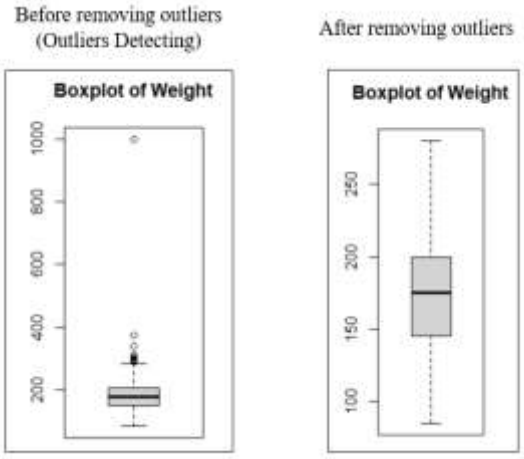

Figure 2. Before & after removing outliers of weight with boxplot

Other than the boxplot, we used the histogram to understand whether there are outliers exist in the dataset easily. The purpose is to have a clear view of outliers that exist in our data.

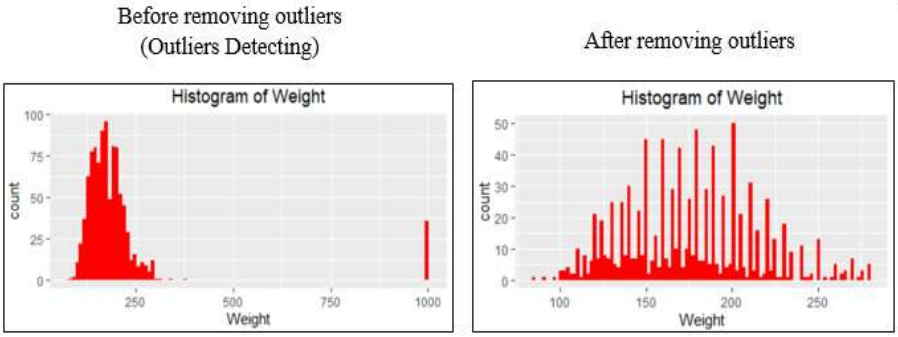

Figure 3. Before & after removing outliers of weight with histogram

Submission: 17 October 2021
Acceptance: 11 March 2022

As for data reduction and relationship measurement among variables, we employ Cramer's V, Correlation Coefficient, and P-value. The Cramer's V and the Correlation Coefficient are two of the data reduction methods that we use in this study. Cramer's V is a value to measure and understand the strength of a relationship between two categorical variables (RPubs, 2020). Cramer's V gives the value between 0 to positive 1. The relationship is stronger when the value is closer to positive 1. When Cramer's V is 1, it means the categorical variables have the perfect association. The main reason for choosing Cramer's V in our datasets is due to many categorical data exist whereby Cramer's V is specialized to discover the correlation ship between categorical data (RPubs, 2020).

As for numerical data in our dataset, we exploited Correlation Coefficient to calculate the co-relationship value. Correlation Coefficient is a statistical concept that aids in the establishment of a link between expected and actual results in a statistical experiment. The correlation coefficient's computed value explains the exactness of the anticipated and actual values. The value of the Correlation Coefficient is always between -1 and +1. If the correlation coefficient is positive, the two variables have a comparable and same relationship. Otherwise, it signifies a difference between the two variables.

P-value calculation will be used to reject the null hypothesis if the value less than the significance level of 0.05. It concludes that the two variables are independent. After the p-value, Cramer's V, and Correlation Coefficient were identified, we distributed the attributes and provide an interpretation of the association. The interpretation of association is divided into three groups i.e. the Moderate, weak, and negligible. The "Moderate" group will only contain the attributes with the Cramer's V or Correlation Coefficient that their value is more than 0.200. The "weak" group will be containing the attributes which value is from 0.100 to 0.200. The last group is "negligible", this group is where the attributes that their value are 0.100 and below. By the time we distributed each attributes to its particular groups, we transfer the moderate and weak groups to other datasets for further use.

Apart from that, we also use mosaic plots to identify the connection between two or more categorical variables. A mosaic plot is a multidimensional version of spine plots, which graphically represent the same data for a single variable. It provides a summary of the data and allows the identification of connections between variables.

Figure 4 below is referred for the following explanation of sleep disorder against depression, anxiety disorder, felt sad or depressed and overall health. The higher amount of the Standardized Residuals is shown as Blue color and the lower amount of the Standardized Residuals is shown as Red color. If a plot has a high amount of Standardized Residual, it means the observed value has a significant relationship with the expected value.

The first comparison in the mosaic plot in figure 4 is between "Sleep Disorder" and the "Depression". It has the highest Cramer's value among all the other categories, which is 0.3276. The more the Cramer's value near the value 1, the more significant the relationship between those two categories. As the showing of the plot, when the patients have depression, they will have a high amount of risk to get sleep disorder according to the datasets.

The second comparison in the mosaic plot in figure 4 is between "Sleep Disorder" and the "Anxiety Disorder". Cramer's value for both is 0.2594. It is considered a rather high value among the other categories. The plot in figure 4 showed a high amount of the Standardized Residuals so it proves the "Sleep Disorder" and the "Anxiety Disorder" have a significant relationship. The plot indicated that if people have an anxiety disorder, they have a higher percentage of a chance of getting a sleep disorder.

Submission: 17 October 2021
Acceptance: 11 March 2022

The third comparison in the mosaic plot in figure 4 is between Sleep Disorder and the respondents of the surveys who have not the feeling of sadness and depression. From viewing the mosaic plot, it is showed people who felt sad or depressed every day or a few days a week have or had a sleep disorder. Consequently, there is a strong relationship between Sleep Disorder and the "How often felt sad or depressed". The Cramer's value obtained by this is 0.2727.

The last comparison in the mosaic plot in figure 4 is observing the relationship between Overall Health and Sleep Disorder. Cramer's value for both is 0.2821. It is obvious between Overall Health and Sleep Disorder have a strong relationship. Although Overall Health and Sleep Disorder have a strong relationship they cannot directly prove that good health will affect people to get sleep disorders. It is because the Standard Residuals have spread through all the attributes in the "Overall Health" category. Hence, the people who have good health may have another factor that causes them to experience Sleep Disorders.

All data involved in the analysis in Figure 4 have been undergone a data reduction process that is depicted in Table 1. Based on the values obtained from Cramer's V, Correlation Coefficient, and P-value, there were 4 attributes identified as moderate and 10 attributes identified as weak whereby others are categorized as negligible. After completed this part, and finalized the most significant attributes from multiple datasets, we merged all significance attributes and come out with the final datasets as depicted in Table 2.

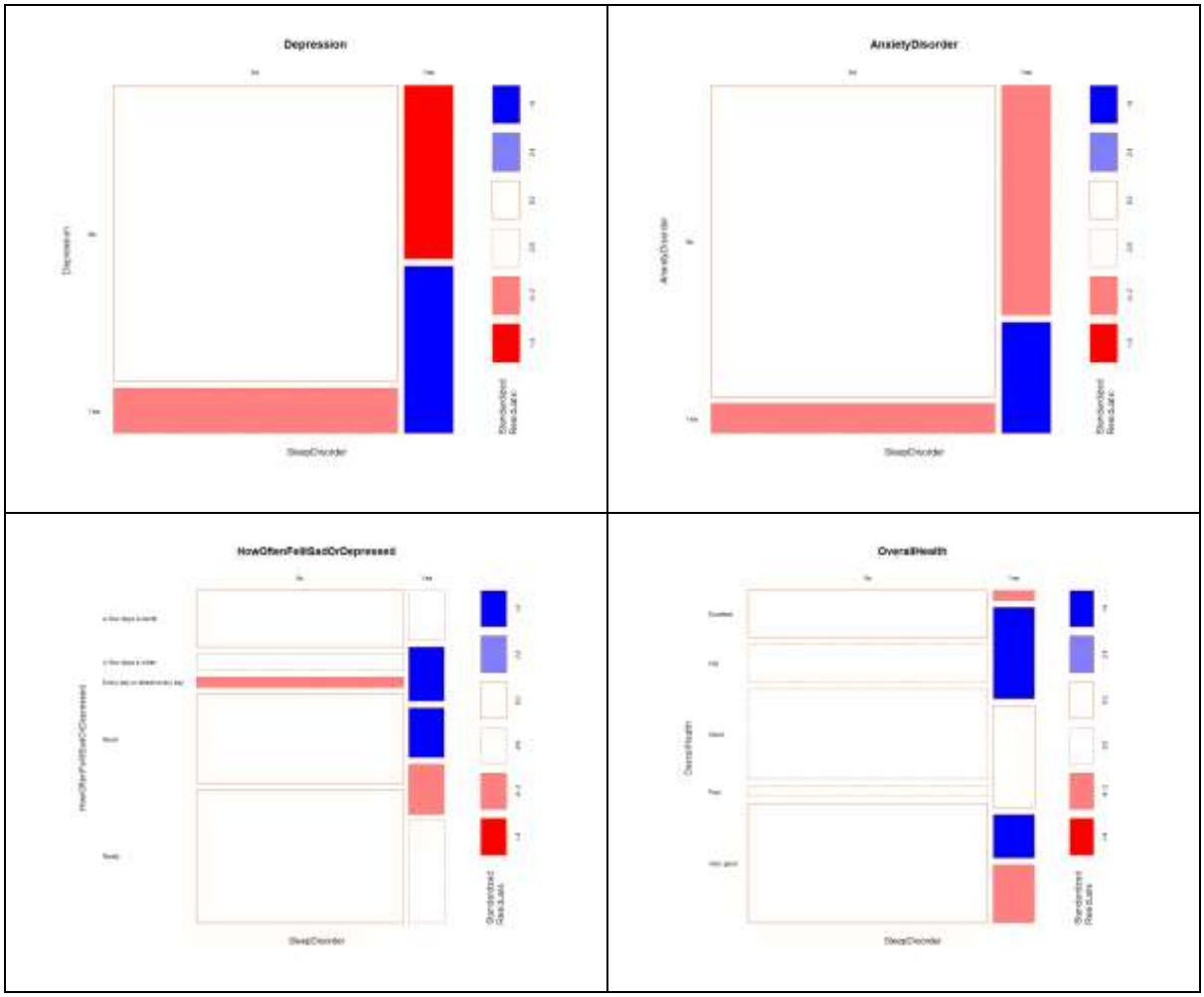

Figure 4. Mosaic Plot for depression, anxiety disorder, felt sad/depressed and overall health

Submission: 17 October 2021
Acceptance: 11 March 2022

Table 1. Data Reduction Process and Attribute Selection

Attribute	p-value	Cramer's V / Correlation Efficient	Interpretation of Association
Depression	2.2e-16	0.3276	Moderate
OverallHealth	2.2e-16	0.2821	Moderate
HowOftenFelltSadOrDepressed	2.712e-15	0.2727	Moderate
AnxietyDisorder	7.183e-16	0.2594	Moderate
HowOftenFeltWorried	5.052e-08	0.1992	Weak
UnpleasentFeelingsInLegs	8.619e-08	0.1963	Weak
Diabetes	4.148e-07	0.1639	Weak
Arthritis	2.376e-06	0.1523	Weak
HeartDisease	4.728e-06	0.1491	Weak
HighBloodPressure	9.208e-06	0.1431	Weak
ChronicIllness	2.69e-05	0.1369	Weak
BreathingDifficulties	0.0004002	0.1158	Weak
ActivityDidBeforeSleep.SurfedTheInternet	0.005997	0.1016	Weak
HaveTwoOrMoreAlcoholicDrinks	0.03498	0.1013	Weak
UsedFunctionsOfCellPhone.Send.ReadOrReceiveTextMessages	0.006997	0.0983	Negligible
ActivityDidBeforeSleep.WatchedAVideoOnYourComputer.Laptop.PhoneOrOtherDeviceThatIsNotATV	0.005497	0.09472	Negligible
Weight	0.003683	0.09175598	Negligible
ActivityDidBeforeSleep.Sent.ReadOrReceivedTextMessages.,	0.003998	0.09024	Negligible
UsedFunctionsOfCellPhone.SurfTheInternet.,	0.006497	0.08653	Negligible
ActivityDidBeforeSleep.DidHomeworkOnTheComputer	0.0004998	0.08557	Negligible
TechUsedBeforeHourToSleep.ComputerOrLaptop	0.002999	0.08488	Negligible
TechUsedBeforeHourToSleep.Cellphone.,	0.005497	0.08456	Negligible
UsedFunctionsOfCellPhone.UseThePhone.sAlarmClock	0.03848	0.07922	Negligible
ActivityDidBeforeSleep.PlayedVideoOrComputerGame.,	0.03198	0.07911	Negligible
ActivityDidBeforeSleep.UsedSocialMedia.	0.01399	0.07589	Negligible
Cancer	0.0296	0.07324	Negligible
ActivityDidBeforeSleep.SentOrReceivedPersonalEmails.,	0.03498	0.0668	Negligible
SleepTime.Weekdays.,	0.02842	0.06930346	Negligible
TechUsedBeforeHourToSleep.PrintedBookOrMagazine.,	0.03598	0.04884	Negligible

Table 2. Data Integration and Finalized Attributes

	Attribute	Data Type	Details
1	OverallHealth	nominal	Overall health of the respondents (Excellent, very good, good, fair, poor)
2	HowOftenFelltSadOrDepressed	nominal	How often the respondents felt sad or depressed (Every day or almost everyday, A few days a month, A few days a week, Rarely, Never)
3	HowOftenFeltWorried	nominal	How often the respondents felt worried (Every day or almost everyday, A few days a month, A few days a week, Rarely, Never)
4	UnpleasentFeelingsInLegs	nominal	How often the respondents have unpleasant feelings in legs (Every day or almost everyday, A few days a month, A few days a week, Rarely, Never)
5	HeartDisease	nominal	Respondents have heart disease (Yes, No)
6	HighBloodPressure	nominal	Respondents have high blood pressure (Yes, No)
7	Diabetes	nominal	Respondents have diabetes (Yes, No)
8	Arthritis	nominal	Respondents have arthritis (Yes, No)
9	BreathingDifficulties	nominal	Respondents have breathing difficulties (Yes, No)
10	Depression	nominal	Respondents have depression (Yes, No)
11	AnxietyDisorder	nominal	Respondents have anxiety disorder (Yes, No)
12	ChronicIllness	nominal	Respondents have chronic illness (Yes, No)
13	ActivityDidBeforeSleep.SurfedTheInternet	nominal	How often the respondents surf Internet before sleep (Every day or almost everyday, A few days a month, A few days a week, Rarely, Never)
14	HaveTwoOrMoreAlcoholicDrinks	nominal	How often the respondents have two or more alcoholic drinks (Every day or almost everyday, A few days a month, A few days a week, Rarely, Never)
15	SleepDisorder	nominal	Class label (No, Yes)

The results of the prediction models are outlined in the following figures. Using Support Vector Machine (SVM) we explore parameter cost and gamma with a variety of values to reduce the misclassifications. The best result was obtained at cost = 100 and gamma = 1 as illustrated in Figure 5. The accuracy level of SVM outperformed the other two methods with 80.28846% accuracy.

Submission: 17 October 2021
Acceptance: 11 March 2022

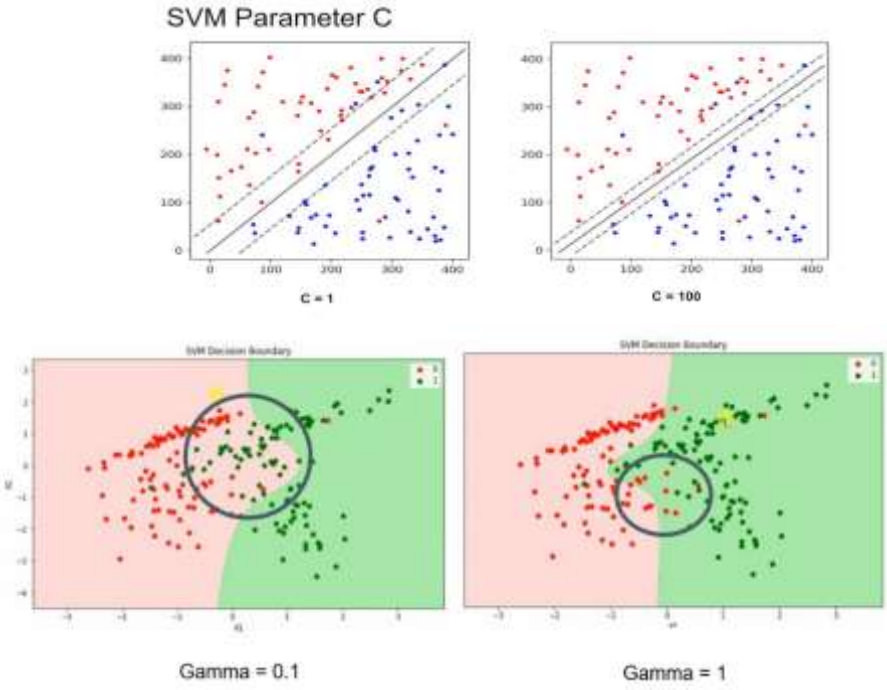

Figure 5. SVM with parameter C and gamma

The results of the Conditional Inference Tree (Ctree) are depicted in figure 6 below. We use a significance test (Permutation test) to select covariates to split and recurs the variables. Three nodes were involved in over 797 observations. The accuracy of Ctree was 66.49937%, the least among others.

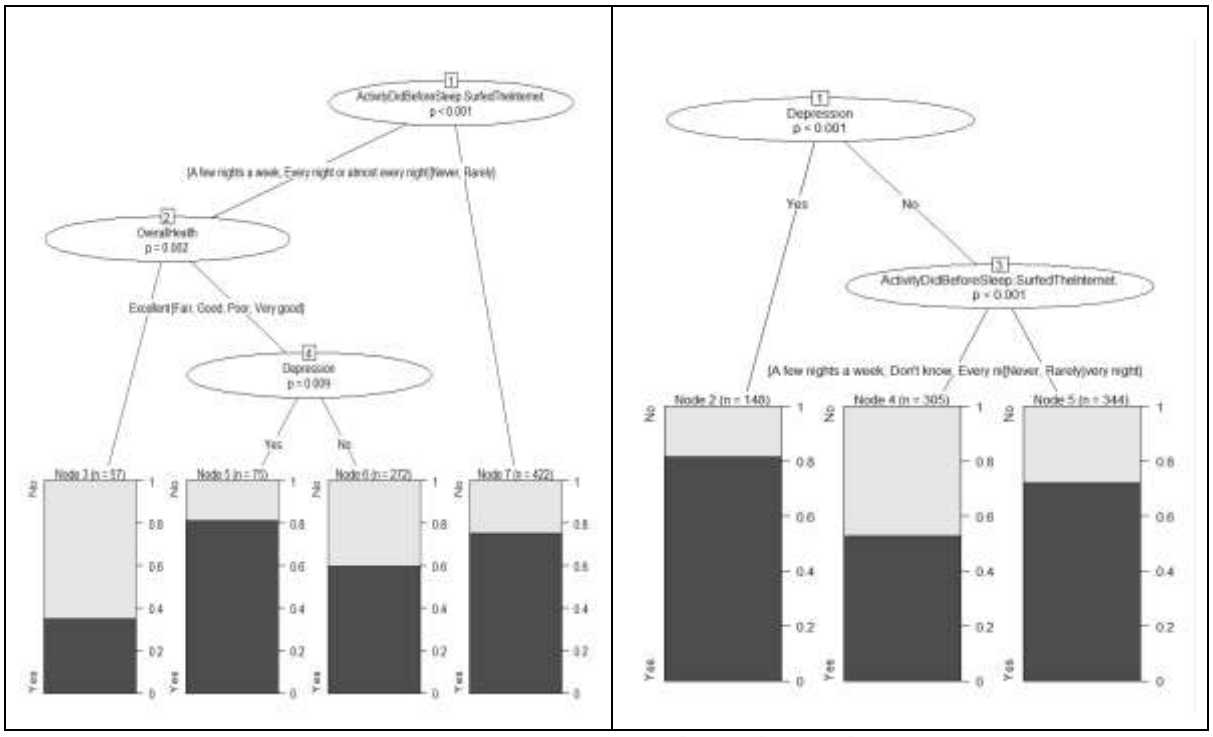

Figure 6. Ctree results with 3 nodes

Submission: 17 October 2021
Acceptance: 11 March 2022

Lastly, the results of Recursive Partitioning (Rpart) are depicted in figure 7. In terms of 100% match comparison between predicted and true value, we were able to achieve 14 as "No" and 131 as "Yes" which drawn a 71.42857% level of accuracy.

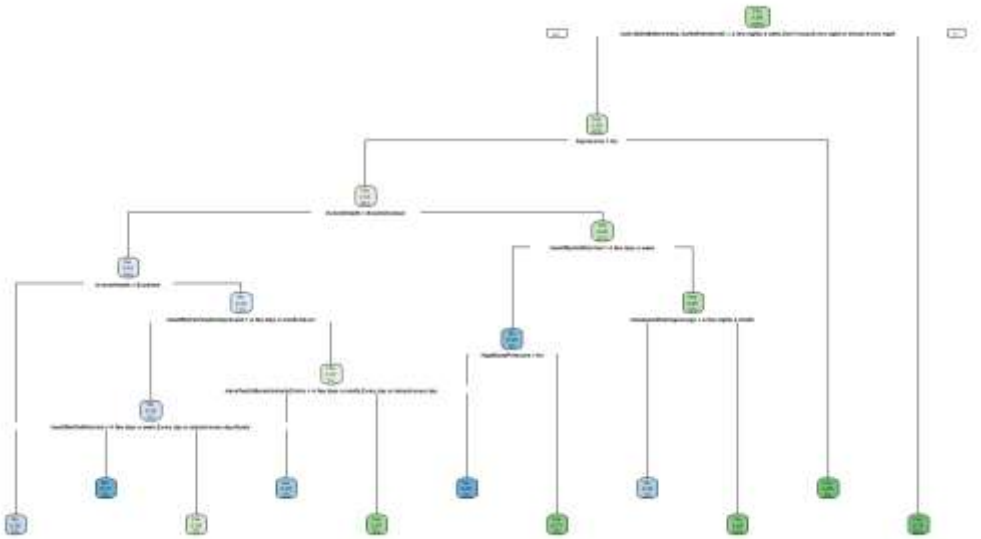

Figure 7. Rpart results

The summary of results over SVM, Ctree, and Rpart is depicted in Table 3. The highest accuracy was achieved by SVM, followed by Rpart and Ctree.

Table 3. Accuracy of machine learning results

Results of SVM	`> # % of correct predictions`
	`> correct = (sum(diag(tab1))/sum(tab1)) * 100 #80.28846% of the predictions were accurate`
	`> correct`
	`[1] 80.28846`
	`>`
	`> #finding the percentage of wrong prediction / misclassfication`
	`> wrong = (1-(sum(diag(tab1))/sum(tab1))) * 100`
	`> wrong`
	`[1] 19.71154`
	`>`

Results of Ctree	`> #finding the correct prediction`
	`> #// 66.49937`
	`> r1 = (sum(diag(tab1))/sum(tab1))*100`
	`> r1`
	`[1] 66.49937`
	`>`
	`> #finding the percentage of wrong prediction / misclassification`
	`> r1 = (1-(sum(diag(tab1))/sum(tab1))) * 100`
	`> r1 #// 33.50063`
	`[1] 33.50063`

Results of Rpart	`> # % of correct predictions`
	`> correct2 = (sum(diag(tab2))/sum(tab2)) * 100 #71.43% of the predictions were accurate`
	`> correct2`
	`[1] 71.42857`
	`>`
	`> #finding the percentage of wrong prediction / misclassfication`
	`> wrong2 = (1-(sum(diag(tab2))/sum(tab2))) * 100`
	`> wrong2`
	`[1] 28.57143`

Submission: 17 October 2021
Acceptance: 11 March 2022

4. Conclusions

The correlation between sleep disorders and different lifestyles has been successfully identified in this study. It showed that are 14 attributes that were found to be significant towards sleep disorder however the 4 most attributes were depression, overall health, felt sad/depressed, and anxiety disorder. The average results of the accuracy of the three predictive models to measure the relationship between sleep disorder and lifestyle is 72.7338 whereby the highest accuracy is sustained by the SVM method. The value makes the SVM worth consideration in predicting new data as well as further improvement or implementation. To enable wider options of predictive models, it is necessary to continue this experiment with other machine learning algorithms to find the strongest performance.

5. REFERENCES

Bhaskar, S., Hemavathy, D., & Prasad, S. , 2016. Prevalence of chronic insomnia in adult patients and its correlation with medical comorbidities. [Online] Available at: https://pubmed.ncbi.nlm.nih.gov/28348990/[Accessed July 2021].

Bjorvatn, B., Grønli, J. and Pallesen, S., 2010. Prevalence of different parasomnias in the general population. Sleep Medicine, 11(10), pp.1031-1034.

Campsen, N. and Buboltz, W., 2017. Lifestyle Factors' Impact on Sleep of College Students. Austin Journal of Sleep Disorders, 4(1).

CentersforDiseaseControlandPrevention, 2020. Drowsy Driving: Asleep at the Wheel. Online] Available at: https://www.cdc.gov/sleep/features/drowsy-driving.html [Accessed July 2021].

ClevelandClinic, 2020. Common Sleep Disorders. [Online] Available at: https://my.clevelandclinic.org/health/articles/11429-common-sleep-disorders [Accessed July 2021].

Donovan, L. and Kapur, V., 2016. Prevalence and Characteristics of Central Compared to Obstructive Sleep Apnea: Analyses from the Sleep Heart Health Study Cohort. Sleep, 39(7), pp.1353-1359.

Patel, D., Steinberg, J., & Patel, P., 2018. Insomnia in the Elderly: A Review. [Online] Available at: https://pubmed.ncbi.nlm.nih.gov/29852897/[Accessed July 2021].

MedlinePlus, 2020. Restless legs syndrome: MedlinePlus Genetics. [Online] Available at: https://medlineplus.gov/genetics/condition/restless-legs-syndrome/ [Accessed July 2021].TechVidvan. 2021. SVM in R for Data Classification using e1071 Package - TechVidvan. [Online]Available at: https://techvidvan.com/tutorials/svm-in-r/#:~:text=An%20SVM%20model%20is%20a%20representation%20of%20the,a%20plane%20%28in%20case%20of%20the%203D%20plane%29 [Accessed July 2021].

Submission: 17 October 2021
Acceptance: 11 March 2022

Topics, H., 2021. Sleep Disorders | MedlinePlus. [Online] Available at: https://medlineplus.gov/sleepdisorders.html [Accessed July 2021].

Submission: 17 October 2021
Acceptance: 11 March 2022