

The Effects of Health and Wellness on the Presence of Sleep Disorders

An approach to predicting sleep disorders using machine learning

Ceara Stewart | IST707 | 9/16/2023

# Introduction

Sleep is one of the most important things to remain healthy. While sleeping, your body has a chance to recover and maintain healthy brain function. And one’s quality and length of sleep can have huge impacts on even small day to day activities. Sleep affects your heart and circulatory system, metabolism, raspatory system, and immune system. But not everyone can obtain good sleep, and sleeping disorders like insomnia and sleep apnea can result in potential health problems. Insomnia is where an individual has trouble falling or staying asleep, while sleep apnea is where the individual experiences pause in breathing during sleep. But sleep doesn’t just influence one’s health, as one’s health and daily habits can influence their sleep, specifically if they have a sleeping disorder.

This paper aims to utilize machine learning models to predict if patients have or do not have a sleeping disorder based on general health and well-being data. The machine learning models that will be used to make predictions include k-Means, EM algorithm, decision trees, and random forests. The goal is to create models with high accuracy that utilize training data of patient health and well-being and their associated sleeping disorder or lack of to predict if test data of patient health and well-being can be associated with having or not having a sleep disorder.

# Data and its Source

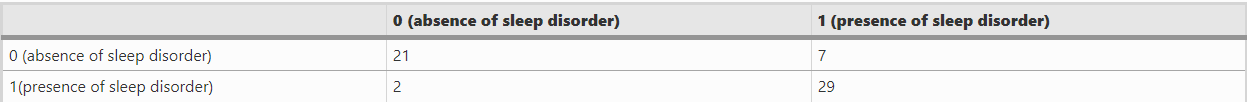
There are two data sets used for this analysis. The first is a data set from Kaggle called the “Sleep Health and Lifestyle Dataset”, that holds a variety of health and well-being variables along with a variable that lists whether the associated participant has insomnia, sleep apnea, or has no sleep disorder. The second data set is self-collected data from colleagues, friends, and family. It was collected completely anonymously and has all the same variables as the data set from Kaggle, except the sleep disorder variable for every participant is listed as “unknown”. The first data set is utilized as a training data set while the second data set is the test data set.

There are twelve relevant variables within the datasets as well as an identification variable. The relevant variables are gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress level, BMI category, blood pressure, heart rate, daily steps, and sleep disorder. Age, sleep duration, sleep quality, physical activity level, stress level, blood pressure, heart rate and daily steps are all quantitative variables while gender, occupation, BMI category and sleep disorder are categorical variables. Whether a participant has a sleep disorder (insomnia, sleep apnea, or none) is predicted from the remaining eleven relevant variables.

# Literature Review

“Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown”, is written by Anbarasi, Jawahar, Cherian and Shreenidhi and this study sought to investigate the different symptoms and side effects that people suffer from in connection to their sleep patterns and uses machine learning to predict the possible relationships that exist between COVID-19 and sleep patterns and disorders. The machine learning methods that this paper uses to come to its conclusions are random forest models and k-Mean clustering techniques.

The random forests model is a step up from the decision tree model which only uses a single tree, while random forest can grow multiple trees. It is used for both classification and regression and overcomes the error due to bias, a major drawback of the decision tree model. In the “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown” journal, random forest is used to build a classification model where the classes were the presence of sleep disorder or the absence of it. The confusion matrix produced by post-training on the data is presented below. There are 21 instances of a true negative, 2 instances of a false negative, 7 instances of a false positive, and 29 instances of a true positive. This model from the journal has high true negative and true positive results and can be considered a good model. Ideally, we want to see a model with the lowest possible false negative and false positive results, and I hope that with the dataset I will be using to conduct my own machine learning research, I can get a smaller percentage of false results as the false negative and false positive percentages are above 8% of the total dataset. The “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown” journal mainly focuses on two variables, anxiety and sleep quality, and their effects on the presence of a sleep disorder. My goal is to create a more ideal random forest model and subsequent confusion matrix using variables such as sleep quality along with ten other variables such as blood pressure and daily steps, and then utilize model tuning methods, to decrease the percentage of false positives and false negatives that are present in the random forest confusion matrix.



Random Forest Confusion Matrix from “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown”

The next machine learning model that the “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown” journal utilizes is the K-means clustering algorithm. K-means clustering is one of the most popular algorithms used for clustering techniques. The algorithms’ goal is to minimize the variance within a cluster while also trying to maintain high variance among the data points of different clusters. The journal uses K-means clustering on participants anxiety and sleep quality scores. The results of the K-mean clustering present that anxiety score does play an effect on whether the participants were classified as either having or not having a sleeping disorder, while sleep quality clustering had less of an effect. The goal of running the k-means clustering on my dataset is to find a model with high accuracy and where more variables can possibly be found to influence whether a participant has a sleeping order or not.

This journal, although provides good insights into the various effects of variables such as occupation, anxiety, and sleep quality on whether a participant does or does not have a sleeping disorder, has a few shortcomings. One shortcoming is that the “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown” journal is that is groups sleep quality with having a sleeping disorder. This should not be assumed as they might be dependent on each other, but low sleep quality scores do not instantly mean an individual has a sleeping disorder and high sleep quality scores do not mean an individual does not have a sleeping disorder. Another shortcoming is that this paper also alludes to the idea that anxiety levels also immediately correlate with having a sleeping disorder. This paper should not assume that a patient’s anxiety and sleep quality scores are directly linked to a sleep disorder and to avoid this shortcoming in my own research, stress/anxiety and sleep quality scores will be solely independent variables used to predict wither an individual has or does not have a sleeping disorder.

# Data Exploration and Data Cleaning

The “Sleep Health and Lifestyle Dataset” was presented as an Excel file that contains thirteen columns labeled: Person ID, Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, and Sleep Disorder. This data set was utilized as a training set and was loaded as an Excel file in R and converted into an ARFF file for use in Weka.

A testing set was created by asking colleagues, classmates, friends, and family members to enter their own data for each column. For the Blood Pressure column, participants either entered their most recent reading from a doctor’s visit or I utilized a blood pressure cuff to take their blood pressure. Heart Rate was filled out utilizing technology such as Apple Watches, which continuously track the users heart rate, or by having the participant count how many beats they felt in ten seconds and multiplying that by six. Daily steps were calculated by having the participants utilize their Apple Health daily step readings and providing the average daily steps taken in a week. The Sleep Disorder column was filled in as “unknown”. All other columns were filled in with self-selected scores or information. The participant’s data was filled in completely anonymously and then re-arranged to add an extra layer to maintain integrity with personal information. This testing data set was also converted into an ARFF file for use in Weka.

For both data sets, the Person ID column was ignored in all models run as it has no effect on any other column and is just an identification column for the participants. It is removed from the datasets in R, and it is ignored when running the models in Weka. It is important to remove this column so it does not group into clusters when running k-Means, or show up as a node when building decision trees or random forests.

For k-Means and decision tree models run within Weka, the data needed to be normalized to ensure that the models were producing the correct outputs and accuracies. For random forests, choosing to normalize is a tuning parameter. Within R, the data was not normalized as regression models can handle the numeric data columns.

k-Means

K-Means was run utilizing Weka. Three clusters were run for the purpose of simplicity, choosing 500 iterations and with a set seed of 10. I also chose to use the Euclidean distance. The number of clusters, iterations, set seed, and distance type were kept constant throughout all algorithms.

To look at how the test data with the unknown sleep disorders compares to how the training data is clustered, a combined dataset was created. This allowed for a comparison of how the unknown sleep disorders were clustered with the known sleep disorders.

EM Algorithm

The EM algorithm focuses on using log-likelihoods to determine the distribution of the latent variables. To keep the different cluster algorithms uniform across all models run, the EM Algorithm was also run using 3 clusters, 500 iterations, and with a set seed of 10.

Similarly to the k-Means algorithm, to test data with the unknown sleep disorders compared to how the training data is clustered, a combined dataset was created. This allowed for a point-blank comparison between clusters on sleep disorders versus other variables.

Decision Trees

Weka uses the J48 as an implementation of the C5.5 algorithm to construct decision trees. The J48 provides many parameters to tune that result in the production of different decision trees. Some tunning of parameters that are conducted within the output and analysis section of this paper, include the “BinarySplit” option within Weka, where a “True” produces a deep tree with two branches at each level, and a “False” produces a wide tree with many branches at each level. The next parameter is “unpruned”, where a “True” grows a tree completely without pruning, and a “False” prunes the tree. The “ConfidenceFactor” parameter can range from 0 to 1 and decides how aggressively the tree is pruned, where smaller values incur more pruning. The “minNumObj” is another parameter that can be tuned that represents the minimum number of examples in the leaf node, with integers ranging from 1 to infinity. The default integer is 2, and it means all the leaves with only one data example are pruned, and increasing this value results in more aggressive pruning.

Before the model was tuned to produce the “best” and most interpretable model, the training ARFF file was loaded. I ensured that the target parameter was set to “sleep disorder” as that is what we are trying to test. The most accurate model was compared to the Excel test data, where the participants data was compared to the decision tree node splits until a sleep disorder category was reached at the end node.

Random Forests

Unlike other algorithms run within Weka, Random Forest do not necessarily need to be discretized before running a model. To keep the result consistent with what was run when looking at decision trees, all models are run using the training set and applying the visualization results to the Excel test data set to predict the participants sleep disorder or lack thereof.

The Random Forest has the numTrees of numIterations parameter available for tuning. For my version of Weka used, the numIterations default is 100. Tuning the number of trees is important as the more trees you add, the more likely the data will be overfit to the model.

# Output and Analysis

k-Means

The k-Means algorithm focuses on the centroid of the clusters which is the gravity center of the cluster, for which we use the mode for nominal data. When looking at the cluster data, we can interpret the values for each column within each cluster as the mode of that column. The cluster mode selected was “Classes to cluster evaluation” and the sleep disorder column was selected.

When comparing the sleep disorder attribute to different health and lifestyle columns using the visualization feature in Weka, you can see how the unknown sleep disorder classification compares to the known sleep disorder classifications. Below is a screenshot of how sleep disorder classification compares to the different quality of sleep scores. Cluster 2 mainly belongs in the “none” sleep disorder category and is associated with quality of sleep scores of 6 and 7. Participants falling within Cluster 3 mainly fall between sleep apnea and insomnia. And cluster 1 mainly belongs in the “none” sleep disorder category with quality of sleep scores mainly belonging to 8 or 9. When looking at the participants with the unknown sleep disorder category, six out of twenty participants are in cluster 3, three belong to cluster 1, and eleven were placed in cluster 2. Based off the clusters these unknown sleep disorder participants were placed in, along with their aligned quality of sleep score, it can be assumed that two of those twenty participants can be classified as having sleep apnea, one being classified as having insomnia, seven being classified as either having sleep apnea or insomnia, and ten being classified as having no sleep disorder.

A screenshot of a computer

Description automatically generated

k-Means Visualization of Sleep Disorder versus Quality of Sleep

EM Algorithm

The expectation maximization algorithm focuses on using log-likelihoods to determine the distribution of the latent variables. It is useful in large and noisy datasets like this one as it estimates the underlying parameters of a probabilistic model. When comparing each sleep disorder category against different variables, the “none” category I mainly made up of clusters 1 and 3. The sleep categories of “insomnia” and “sleep apnea” are mainly composed of points from cluster 2.

A screenshot of a computer

Description automatically generated An interesting comparison can be found from the EM model. The first looks at sleep disorder versus heart rate, shown below. For sleep disorder is categorized as “none”, no points are found above a heart rate of 77, while “sleep apnea” and “insomnia” have multiple points above a heart rate of 77. For the “unknown” sleep disorder, there are 12 points belonging to cluster 1, 6 points belonging to cluster 2, and 2 points belonging to cluster 3. You can also notice that the points belonging to cluster 2, which is the main cluster split between sleep apnea and insomnia, are all located at higher heart rate values. This could possibly indicate a connection to increased heart rate to the patient having a sleep disorder. So, one could assume that the 8 “unknown” participants having a heart rate above 77 could either have sleep insomnia or sleep apnea. This closely aligns with what was found using the k-Means algorithm as that algorithm assumes 10 “unknown” participants have either sleep apnea or insomnia. When looking at it based on a cluster standpoint, 12 participants belong to cluster 1 and can be assumed, based on where the majority of cluster 1 falls, to have no sleep disorder. While six participants belong to cluster 2 and can be assumed to have either sleep apnea or insomnia, and 2 participants were placed in cluster 3, and could be assumed to not have a sleep disorder.

EM Visualization of Sleep Disorder versus Heart Rate

Decision Trees

I first only ran the J48 model on the training set alone to first try and obtain the most accurate model. The first J48 model I ran kept the preset parameters as is, except for the “binarySplits”. The “binarySplits” parameter was set to “True”, the “confidenceFactor” was set to 0.25, the “minNumObj” was set to 2, and the “unpruned” parameter was set to “False”. The test parameter was also set to “sleep disorder” as that is the attribute from the data set that we are trying to predict. This first model decision tree produced an accuracy of 92.5134%.

I next tuned the “confidenceFactor” to see if a more accurate model could be produced. All other parameters from the first model run were kept constant. Decreasing the confidence factor to 0.1 has no effect on the accuracy of the model. And increasing the confidence factor to 0.3 then 0.5 also had no effect on the accuracy of the model.

The next parameter I tried tuning was the “minNumObj”. I first set the integer to 1. This produced an accuracy of 93.0481%. When the “minNumObj” was increased to three, the model accuracy dropped to 91.9786%. Increasing the “minNumObj” to 4 dropped the accuracy further to 90.107%.

A diagram of a network

Description automatically generated While keeping the “confidenceFactor” at 0.25, and the “minNumObj”at 1, I switched the “unpruned” parameter to “True”. This produced a decision tree with an accuracy of 93.3155%. Although the decision tree is larger and has more node’s, increased accuracy is preferred. This decision tree is shown below.

Decision Tree Visualization of most Accurate Tree

I utilized the tree on the Excel version of the test data to see if each participant could be classified with a sleep disorder. Following the nodes of the tree and their split value, I found that eleven out of twenty participants were classified as having no sleeping disorder, two were classified having insomnia, and seven were classified as having sleep apnea. These results compare similarly with what was found running k-means. For no sleep disorder, k-Means produced a number of 10 while the decision trees model produced a number of 11. The k-Means model classified one participant with definitely having insomnia and two participants definitely having sleep apnea and seven participants having either sleep apnea or insomnia, while the decision tree model classified two participants with having insomnia and seven having sleep apnea.

Random Forests

I first ran the random forest model with its default numIterations set to 100. This default parameter value results in an accuracy of 93.3155%. I then tuned the parameters and set the numIterations to 10, 25, 50, 75 and 110 to see if a more accurate model could be found. The numIterations values produced were all 93.3155%, so changing the number of iterations had no effect on the model accuracy.

Thus, the best random forest model has a number of iterations of 10, as decreasing the number of iterations ensures that the model is not overfitting the data as this would be a problem when running the same model on the test dataset.

A screenshot of a computer

Description automatically generated

We can also interpret the confusion matrix produced by this random forest model. For sleep disorder equals none, the true positive would be 210 while the false negative would be the sum of the remaining values in that row, so 9. The false negative would be the sum of the remaining values in the column, so 11. And the true negative would be the sum of values of all columns and rows except the values of that class, so 144. For the sleep disorder equal to sleep apnea, the true positive would be 70. The false negative would be 8. The false positive would be 7. And the true negative would be 289. And for sleep disorder equal to insomnia, the true positive would be 69. The false negative is 8. The false positive is 7, and the true negative would be 290. You naturally want the false positive and false negative values to be as small as possible. For all three sleep disorder classifications, the false positive and false negative values are small, less than 6% of the dataset, and thus affirm the model accuracy.

# Conclusion

The purpose of this paper was to see if machine learning algorithms could be used to see if a group of participants could be classified with either having no sleep disorder, insomnia, or sleep apnea from a training set of known participants with sleep disorders. The machine learning algorithms of k-Means, expectation maximization, decision trees and random forests were used to test this prediction ability.

For k-Means, when looking at sleep disorder versus quality of sleep, 2 of the “unknown” participants were classified as having insomnia, seven were classified with having either sleep apnea or insomnia, and ten were classified with having no sleep disorder. The EM algorithm, when looking at sleep disorder versus heart rate, two approaches could be used. The first looked at where most of each cluster fell and comparing what clusters each point for the “unknown” sleep disorder belongs to. This would then assume that 12 participants, belonging to cluster 1, have no sleep disorder. That six participants belonging to cluster 2 have either sleep apnea or insomnia. And 2 participants, belonging to cluster 3, also do not have a sleep disorder. The second approach would be to look at where the cluster points fall for each sleep disorder in relation to heart rate. Since no cluster points for no sleep disorder go above a heart rate of 77, 12 participants have no sleep disorder while 8 could be assumed as having either insomnia or sleep apnea. The random forest model lets one look at a confusion matrix and can be seen to have very low percentages of false positives and false negatives. The decision tree model, and its visualization capabilities, lets us follow tree slits and nodes of just the training data and compare it to the test data. This led to the conclusion that 11 participants have no sleep disorder, two have insomnia, and seven have sleep apnea.

Overall, each of these four different machine learning models have similar results when classifying which of twenty participants with an unknown sleep disorder have sleep apnea, insomnia, or no sleep disorder. But out of the four models, random forests were the best model for this data set. Random Forests generally produce high accuracy outputs and are less sensitive to overfitting compared to decision trees. Although k-Means, EM and decision trees allowed for a visual output of the model results, k-Means are sensitive to outliers, EM is sensitive to initialization, and decision trees are prone to overfitting without proper tuning. Any health and lifestyle data are generally filled with outliers, as someone with a blood pressure reading of not 120/80 doesn’t necessarily have a sleep disorder, or someone with perfectly average and normal data could still have a sleep disorder. This data set was also large with close to 400 participants and overfitting could provide incorrect predictions.

Compared to what was presented in the literature review paper “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown”, the results presented in this paper succeeds and keep sleep disorder as a dependent variable and quality of sleep and stress level as independent variables. This bettered the k-Means output compared to that of the literature review paper. This paper also succeeds in lowering the false positives and false negatives as percentages of the data compared to that of the “Machine Learning Approach for anxiety and sleep disorders analysis during COVID-19 lockdown”. But to make the results of this paper even more accurate for future use, the sample size for the training data could and should be increased to include even more participants. This might decrease the effect that outliers have on the data and result in better running of the machine learning algorithms moved.

# Sources

<https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>

<https://link.springer.com/article/10.1007/s12553-022-00674-7#Sec6>