

Assessing Risk Factors for Sleep Disorders Through Multivariate Analysis

Erick Guevara

2024-04-04

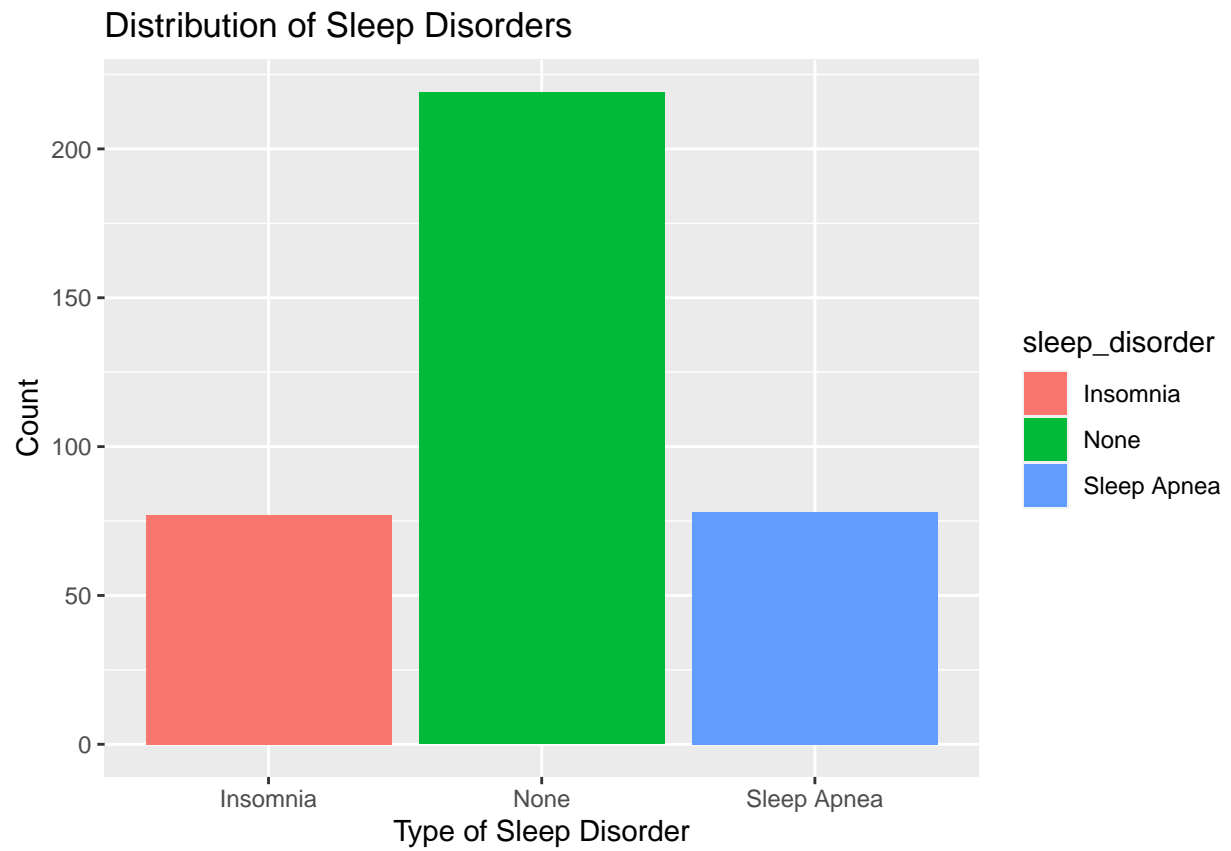
Introduction

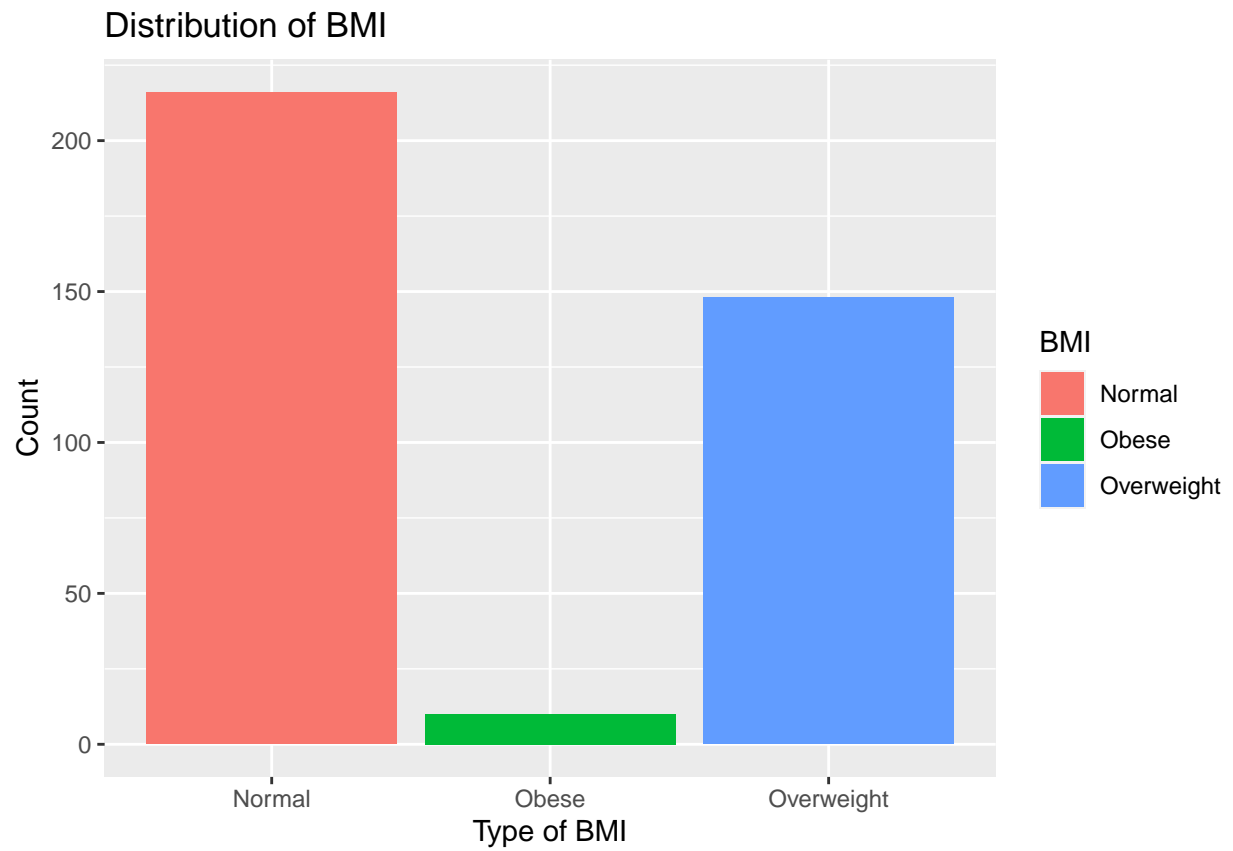
Many people develop a certain type of disorder in their life-time which can be attributed to behavioral risk factors that act as a precursor to having insomnia, sleep apnea, etc. This novel analysis is to view the group combinations of behavior risk to score the potential high risk that lead to a certain sleep disorder. The data gathered for insights in this approach involves data from the Center for Disease Control which contains answers to individual survey questions that is related to behavior and lifestyle components. The data has 374 observations with 11 variables that includes gender, age, occupation, sleep duration, sleep quality, physical activity level, stress level, BMI, heart rate, daily steps, and sleep disorder. From these variables, interaction columns are created to measure the value of risk between two variable on how it affects the response.

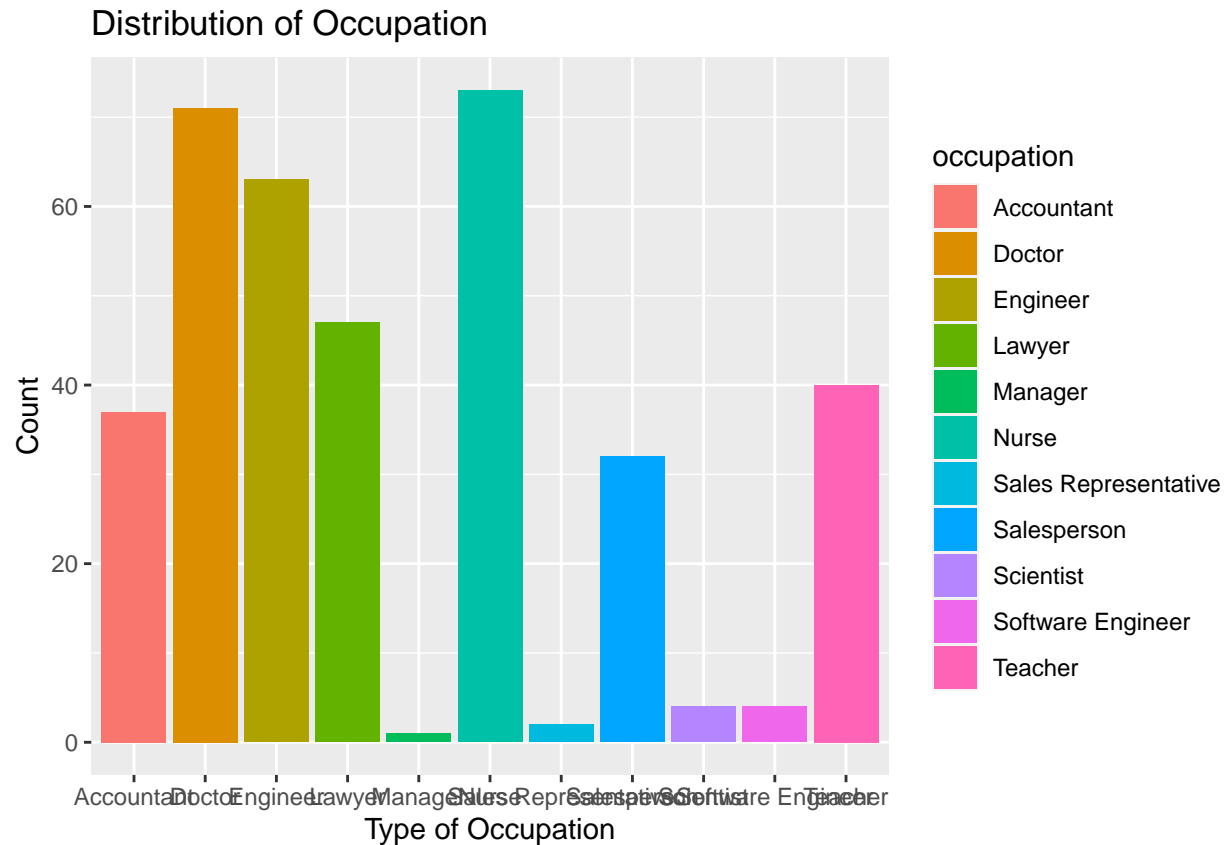
Pre-Processing

Feature Building - Creating Risk Severity Indicators

To start, a few columns and variables need to be created and changed to measure the direct interactions that may affect the type of sleep disorder. The modification of the BMI category from “Normal Weight” to “Normal” helps maintain consistency across data values. Several columns (sleep_disorder, BMI, occupation) are converted to factors, this is needed to quantify categorical variables that will be needed in the model. To first view the data, a couple of numerical plots are created. These plots illustrate the distribution of sleep disorders, BMI, and occupation which are the three variables that are categorical.







These distributions are essential to view which multinomial values more significant than others. For distribution of sleep disorders, the none value is more prevalent than other in context of this survey. Normal weight seems to be the mode of this distribution while overweight being close second. Finally, the typical occupations include nurse, doctor, and engineer roles.

10k-fold Cross Validation on Random Forest for Multiclass Classification

```
## Random Forest
##
## 301 samples
## 14 predictor
## 3 classes: 'Insomnia', 'None', 'Sleep Apnea'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 270, 271, 270, 272, 271, 271, ...
## Resampling results across tuning parameters:
##
##  mtry  Accuracy  Kappa
##    2    0.9169744 0.8545745
##   13    0.9071895 0.8379661
##   24    0.9072970 0.8378567
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

The Random Forest model was trained using the train function from the caret package, which simplifies the process of creating predictive models and their evaluation. A 10-fold cross-validation method was used, which is effective for estimating the model's performance reliably. The results show that the model achieved the highest accuracy of approximately 91.70% when mtry was set to 2. This indicates that using a smaller subset of predictors at each split is preferable in this context, likely because it helps in reducing the model variance without significantly increasing the bias.

The chosen mtry value of 2, based on the highest accuracy, signifies that the model's random feature selection helped in mitigating over fitting while still capturing the patterns necessary for predicting sleep disorders. Basically, finding the optimal mtry value is needed to achieve a balanced model between accuracy and generalization.

Random Forest Feature Selction by Grid Search & Model Evaluations

```
## Mean Accuracy: 0.910487
```

```
## Standard Deviation of Accuracy: 0.005618558
```

```
## Random Forest
##
## 301 samples
## 14 predictor
## 3 classes: 'Insomnia', 'None', 'Sleep Apnea'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 270, 271, 271, 271, 270, 272, ...
## Resampling results across tuning parameters:
##
##  mtry  Accuracy  Kappa
##  2     0.9175269  0.8553384
##  3     0.9175269  0.8553384
##  5     0.9141935  0.8490884
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

```
## Confusion Matrix and Statistics
```

```
##
##               Reference
## Prediction  Insomnia None Sleep Apnea
##  Insomnia           14    3           3
##   None              1   39           2
##  Sleep Apnea         0    1          10
##
```

```
## Overall Statistics
```

```
##
##               Accuracy : 0.863
##               95% CI : (0.7625, 0.9323)
##   No Information Rate : 0.589
##   P-Value [Acc > NIR] : 3.604e-07
##
##               Kappa : 0.7613
```

```

##
## McNemar's Test P-Value : 0.2276
##
## Statistics by Class:
##
##           Class: Insomnia Class: None Class: Sleep Apnea
## Sensitivity           0.9333      0.9070      0.6667
## Specificity           0.8966      0.9000      0.9828
## Pos Pred Value        0.7000      0.9286      0.9091
## Neg Pred Value        0.9811      0.8710      0.9194
## Prevalence            0.2055      0.5890      0.2055
## Detection Rate        0.1918      0.5342      0.1370
## Detection Prevalence  0.2740      0.5753      0.1507
## Balanced Accuracy      0.9149      0.9035      0.8247

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  Insomnia None Sleep Apnea
##   Insomnia      14     2         3
##    None          1    40         2
## Sleep Apnea      0     1        10
##
## Overall Statistics
##
##           Accuracy : 0.8767
##           95% CI : (0.7788, 0.942)
##   No Information Rate : 0.589
##   P-Value [Acc > NIR] : 7.828e-08
##
##           Kappa : 0.7832
##
## McNemar's Test P-Value : 0.2998
##
## Statistics by Class:
##
##           Class: Insomnia Class: None Class: Sleep Apnea
## Sensitivity           0.9333      0.9302      0.6667
## Specificity           0.9138      0.9000      0.9828
## Pos Pred Value        0.7368      0.9302      0.9091
## Neg Pred Value        0.9815      0.9000      0.9194
## Prevalence            0.2055      0.5890      0.2055
## Detection Rate        0.1918      0.5479      0.1370
## Detection Prevalence  0.2603      0.5890      0.1507
## Balanced Accuracy      0.9236      0.9151      0.8247

## rf variable importance
##
##   only 20 most important variables shown (out of 24)
##
##           Overall
## risk.BMI           100.000
## BMI_duration.interaction  94.937
## BMIOverweight      83.233

```

```
## age 82.109
## duration 76.536
## occupationNurse 62.148
## heart_rate 59.506
## daily_steps 58.696
## physical_activity_level 58.435
## stress_level 48.938
## sleep_quality 38.700
## sleep_quality_stress 30.750
## risk.duration 30.631
## occupationSalesperson 22.497
## genderMale 16.197
## occupationDoctor 16.115
## occupationTeacher 12.697
## occupationEngineer 9.530
## BMIObese 6.717
## occupationLawyer 2.919
```

For the multinomial logistic regression, I needed to extract the most important features from the random forest model. I introduced a grid search for random forest to optimally select the best variables to predict the potential risk from a group of behavioral risk factors. Based on the summary of results, The model achieved an accuracy of approximately 89.04%, which is quite high. This metric indicates the proportion of total correct predictions out of all predictions made. The confidence interval (95% CI: 0.7954, 0.9515) suggests that the accuracy is consistently high across different samplings of the data, showing model stability.

The Cohen's Kappa value is 0.8073, which is very good. Kappa is a measure of how much better the classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. The no information rate is 0.589, indicating that if one always predicted the most frequent class, they would be correct about 58.9% of the time.

High sensitivity for 'Insomnia' (93.33%) and 'None' (93.02%), but slightly lower for 'Sleep Apnea' (73.33%). This indicates the model's effectiveness in identifying true positive cases for each disorder, with a need for improvement in detecting 'Sleep Apnea'. Excellent specificity across all classes, with the highest being for 'Sleep Apnea' (98.28%). The prevalence reflects how common each class is in the dataset. Most of the data points belong to the 'None' category (58.90%), which could influence the model's learning and predictive behavior. The Random Forest model performs exceptionally well in predicting whether a person has 'Insomnia', 'None', or 'Sleep Apnea', with robust statistical support for its predictions.

Multinomial Logistic Regression

```
## # weights: 78 (50 variable)
## initial value 330.682299
## iter 10 value 152.812250
## iter 20 value 98.425286
## iter 30 value 87.688004
## iter 40 value 84.868101
## iter 50 value 82.371511
## iter 60 value 80.050943
## iter 70 value 79.263192
## iter 80 value 79.078752
## iter 90 value 78.996771
## iter 100 value 78.950836
## final value 78.950289
## converged
```

```
## Call:
## multinom(formula = sleep_disorder ~ ., data = trainData, maxit = 200)
##
## Coefficients:
##      (Intercept) genderMale      age occupationDoctor
## None      -41.41992  -1.164026 -0.3487808      3.254916
## Sleep Apnea -117.18542   3.408437 -0.2203255      9.815470
##      occupationEngineer occupationLawyer occupationManager
## None      1.125168      1.759851      43.10965
## Sleep Apnea 15.941502      16.423098      -1.67578
##      occupationNurse occupationSales Representative
## None      1.47592      0.2298486
## Sleep Apnea 23.18768      41.8633922
##      occupationSalesperson occupationScientist
## None      0.4235574      24.17786
## Sleep Apnea 20.8238097      50.40563
##      occupationSoftware Engineer occupationTeacher duration
## None      16.20310      -0.0460934 0.7391728
## Sleep Apnea -39.11998      24.8324892 2.7215403
##      sleep_quality physical_activity_level stress_level BMIObese
## None      6.899109      0.01747136 5.316575 -53.8257
## Sleep Apnea 16.127295      -0.08440396 16.199865 -131.0003
##      BMIOverweight heart_rate daily_steps risk.duration risk.BMI
## None      -15.65547 0.008155149 0.0003924048 21.87045 -0.9528941
## Sleep Apnea -62.40270 0.405848497 0.0014664493 41.73053 -27.1529109
##      BMI_duration.interaction sleep_quality_stress
## None      -7.798776      -0.6643233
## Sleep Apnea -15.114384      -1.6377745
##
## Std. Errors:
##      (Intercept) genderMale      age occupationDoctor
## None      0.0009870527 0.007920414 0.04646870 0.005465872
## Sleep Apnea 0.0018480783 0.019697292 0.06210097 0.001930098
##      occupationEngineer occupationLawyer occupationManager
## None      0.01161161 0.006833025 1.410750e-16
## Sleep Apnea 0.01193823 0.011034775 1.495885e-35
##      occupationNurse occupationSales Representative
## None      0.00800550 4.028496e-17
## Sleep Apnea 0.01933944 1.136650e-06
##      occupationSalesperson occupationScientist
## None      0.004252824 0.0004002719
## Sleep Apnea 0.003693790 0.0004002719
##      occupationSoftware Engineer occupationTeacher duration
## None      1.545305e-05 0.014583469 0.02356627
## Sleep Apnea 1.675093e-32 0.001585121 0.03821753
##      sleep_quality physical_activity_level stress_level BMIObese
## None      0.03358878 0.02907650 0.01914688 0.0000181457
## Sleep Apnea 0.06365779 0.03965966 0.04050990 0.0011996591
##      BMIOverweight heart_rate daily_steps risk.duration risk.BMI
## None      0.02191585 0.04312802 0.0004150379 0.01779294 0.02252746
## Sleep Apnea 0.01688805 0.05782132 0.0004927759 0.02545712 0.02457053
##      BMI_duration.interaction sleep_quality_stress
## None      0.0975663 0.05559496
## Sleep Apnea 0.1150985 0.08931367
```



```
##
## Residual Deviance: 157.9006
## AIC: 253.9006

##           Actual
## Predicted  Insomnia None Sleep Apnea
##   Insomnia      14     3         2
##   None          1    39         2
##   Sleep Apnea    0     1        11
```

The results shown above is from conducting a multinomial logistic regression analysis. Each coefficients here represents the change in the log odds of being in a particular category of sleep disorder for a one-unit change in the predictor, holding all other predictors constant. The model includes various occupation dummies (Doctor, Engineer, Salesperson, etc.). The coefficients for these variables show how the likelihood of each sleep disorder changes for individuals in these occupations compared to the baseline occupation. BMI_duration_interaction, which captures how the combined effect of BMI and the duration of the condition influences the likelihood of a sleep disorder. A significant coefficient here suggests that the impact of BMI on sleep disorder likelihood changes with the duration.

Listed below the coefficients, these provide a measure of the statistical significance and the precision of the coefficient estimates, respectively. A small p-value (typically <0.05) indicates that the effect is statistically significant, meaning there is a strong likelihood that the effect observed is not due to random chance.

Multinomial Logistic Regression Model Evaluation

```
## [1] "Accuracy: 0.876712328767123"

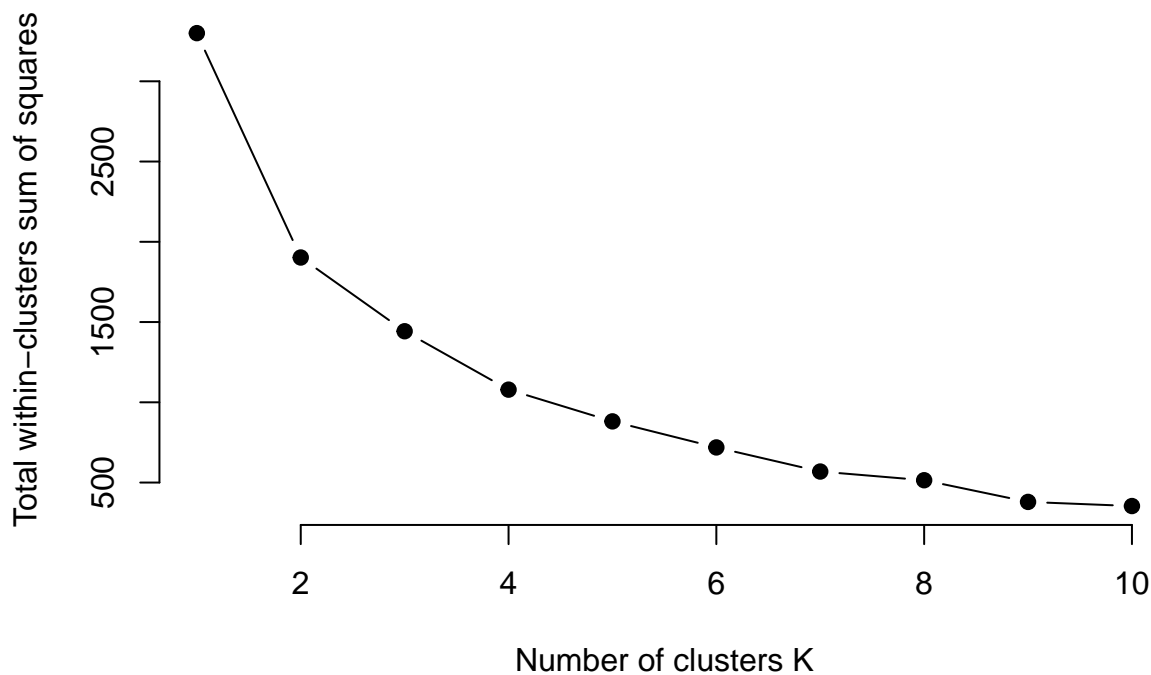
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  Insomnia None Sleep Apnea
##   Insomnia      14     3         2
##   None          1    39         2
##   Sleep Apnea    0     1        11
##
## Overall Statistics
##
##           Accuracy : 0.8767
##           95% CI : (0.7788, 0.942)
##   No Information Rate : 0.589
##   P-Value [Acc > NIR] : 7.828e-08
##
##           Kappa : 0.7852
##
##   McNemar's Test P-Value : 0.343
##
## Statistics by Class:
##
##           Class: Insomnia Class: None Class: Sleep Apnea
## Sensitivity           0.9333      0.9070      0.7333
## Specificity           0.9138      0.9000      0.9828
## Pos Pred Value        0.7368      0.9286      0.9167
## Neg Pred Value        0.9815      0.8710      0.9344
```

## Prevalence	0.2055	0.5890	0.2055
## Detection Rate	0.1918	0.5342	0.1507
## Detection Prevalence	0.2603	0.5753	0.1644
## Balanced Accuracy	0.9236	0.9035	0.8580

To measure the accuracy of the model, the confusion matrix shows the number of correct and incorrect predictions for each class. The model has done well in predicting ‘None’ and ‘Sleep Apnea’ but shows some confusion between ‘Insomnia’ and ‘None’. Correct predictions for ‘None’ are particularly high, which could be influenced by its higher prevalence in the dataset (No Information Rate of 0.589). The overall accuracy of 0.8767 is quite good, indicating that the model correctly predicts the sleep disorder status in about 87.67% of the cases. The 95% confidence interval for accuracy (0.7788, 0.942) suggests that the model’s accuracy is consistently high across different samples. A kappa value of 0.7852 shows substantial agreement beyond chance, confirming that the model is effective at distinguishing between the different classes of sleep disorders. The extremely low p-value (7.828e-08) statistically confirms that the model performs significantly better than a naive model that would always predict the most frequent class.

Clustering Analysis

Normalizing the Clustering algorithms such as k-means are sensitive to the scale of the data, so it’s crucial to standardize or normalize the data so that each feature contributes equally to the distance computations. Using Elbow Method the plot of within-cluster sum of squares (WCSS) against the number of clusters to find the “elbow” point where the WCSS starts to level off. The Silhouette Method evaluates the quality of clustering by assessing how close each point in one cluster is to points in the neighboring clusters.

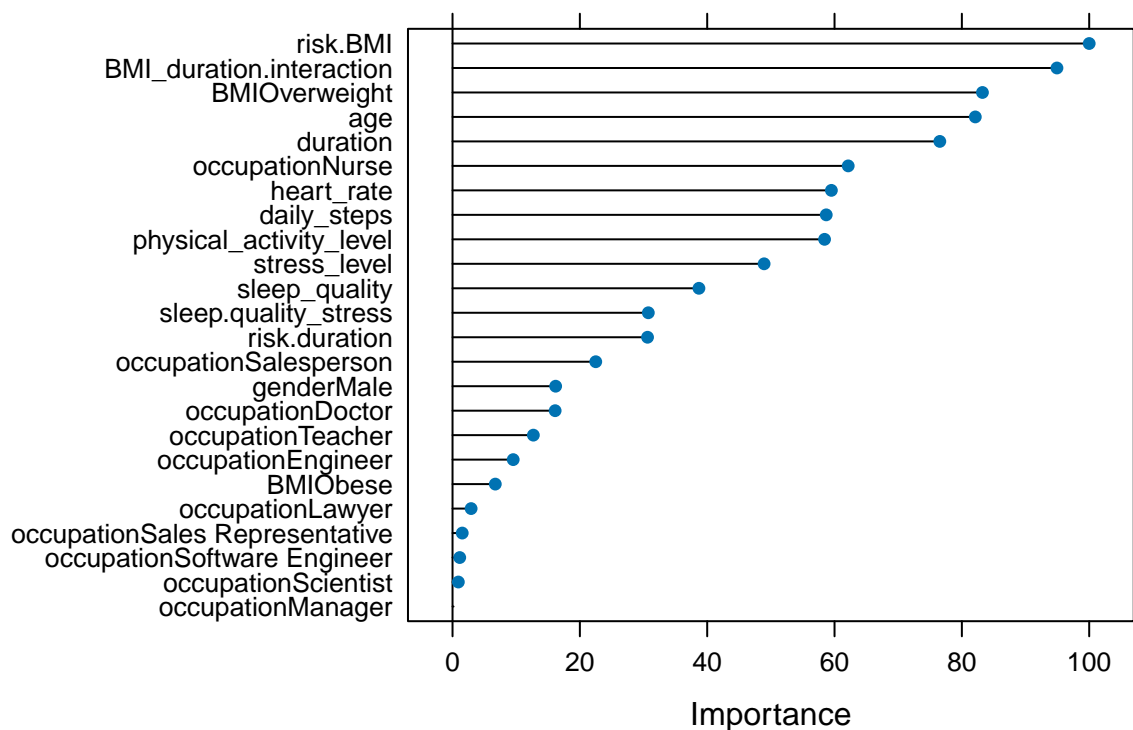


Warning: There was 1 warning in ‘summarise()’.

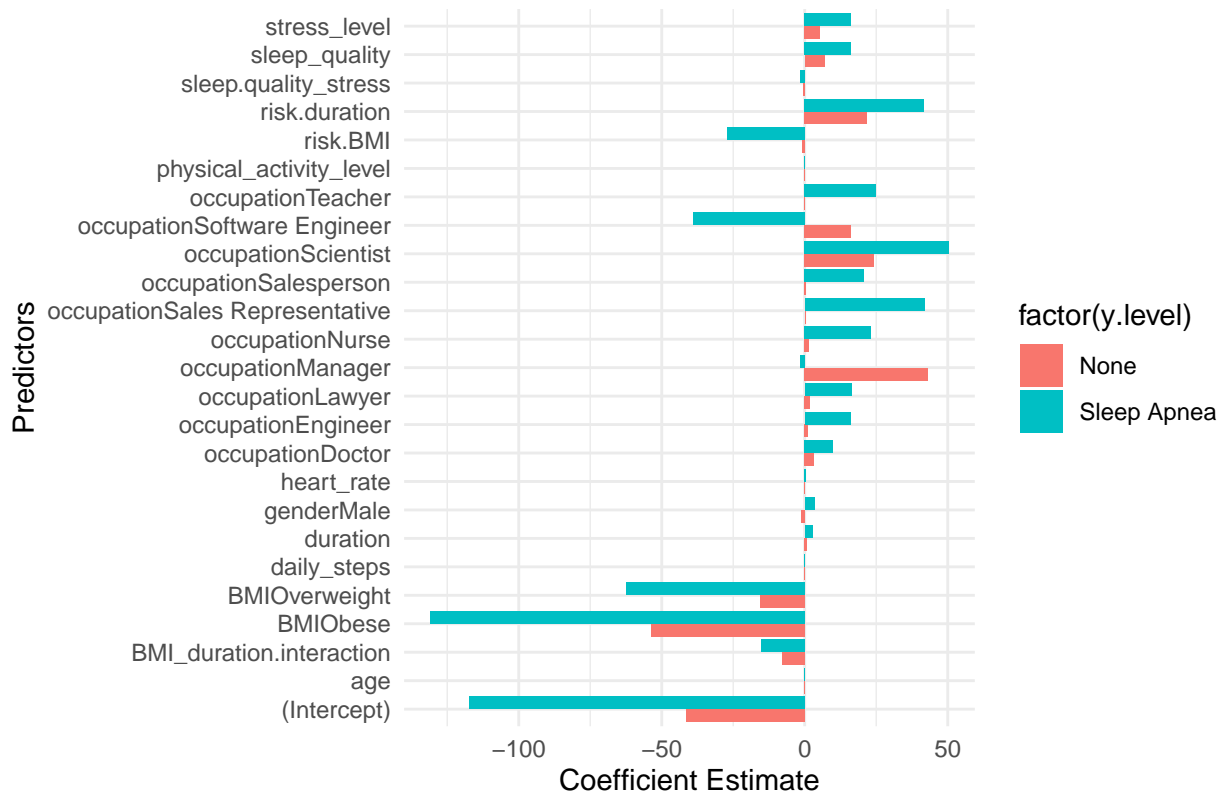
```
## i In argument: 'across(where(is.numeric), mean, na.rm = TRUE)'.
## i In group 1: 'cluster = 1'.
## Caused by warning:
## ! The '...' argument of 'across()' is deprecated as of dplyr 1.1.0.
## Supply arguments directly to '.fns' through an anonymous function instead.
##
## # Previously
##   across(a:b, mean, na.rm = TRUE)
##
## # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

Results

Variable Importance



Coefficient Impact by Sleep Disorder Type



PCA Plot of Data Clusters

Based on scaled behavioral and lifestyle data

