Compare the effects of Random forest and KNN Classification on Sleep Data Set

Eric Wu

2023-09-10

Background

The dataset “Sleep health and lifestyle” is from Kaggle.com, an online community with enriched datasets from various contributors for data analysis and machine learning projects. The primary owner of the “housing” data set is Laksika Tharmalingam.

Figure 1: (Cited from https://opendatascience.com/10-tips-to-get-started-with-kaggle/)

Skill Set Objectives

- KNN Classification on multi-class variable
- Cross Validation
- Data Visualization
- Random Forest on multi-class variable
- Out Of Bag (OOB) Error

Dataset

The “Sleep health and lifestyle” dataset consists of data from 374 individuals with 13 variables. The goal of the dataset is to predict the type of sleep disorders based on all other factors, with an emphasis on sleep and exercise routines. Some of the variables included are sleep duration, quality of sleep, physical activity level, stress level, BMI categories, blood pressure, heart rate, etc. In addition, the dataset also has 0 null values.

Table 1: Basic statistics of quantitative variables

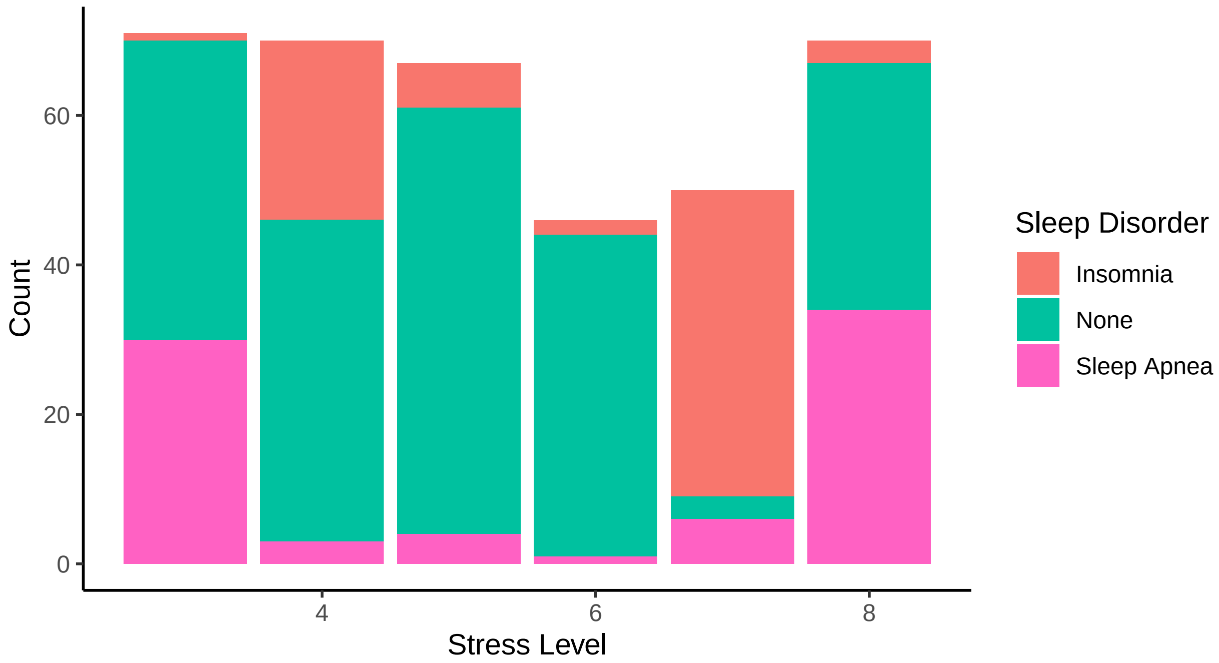
	vars	n	mean	sd	min	max	range	se
Age	1	374	42.18	8.67	27.0	59.0	32.0	0.45
Sleep.Duration	2	374	7.13	0.80	5.8	8.5	2.7	0.04
Quality.of.Sleep	3	374	7.31	1.20	4.0	9.0	5.0	0.06
Physical.Activity.Level	4	374	59.17	20.83	30.0	90.0	60.0	1.08
Stress.Level	5	374	5.39	1.77	3.0	8.0	5.0	0.09
Heart.Rate	6	374	70.17	4.14	65.0	86.0	21.0	0.21
Daily.Steps	7	374	6816.84	1617.92	3000.0	10000.0	7000.0	83.66
Systolic	8	374	128.55	7.75	115.0	142.0	27.0	0.40
Diastolic	9	374	84.65	6.16	75.0	95.0	20.0	0.32

Table 1 demonstrates some sample statistics for the “sleep, health, and lifestyle” dataset, and some of the findings are very interesting. For example, the average value of the quality of sleep is 7.31 (out of 10), which means a majority of people believe they are having quality sleep although only 58.56% don’t have a sleep disorder, this means that 41.44% of people who have a sleep disorder doesn’t seem to be bothered by the sleep disorder. which suggest the importance of another measurement regarding the severeness of the sleep disorder, although it can be estimated with other variables, or suggest that people filling out the survey are being biased about their quality of sleep.

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis section will focus on finding the relationships between Sleep Disorder and other important factors.

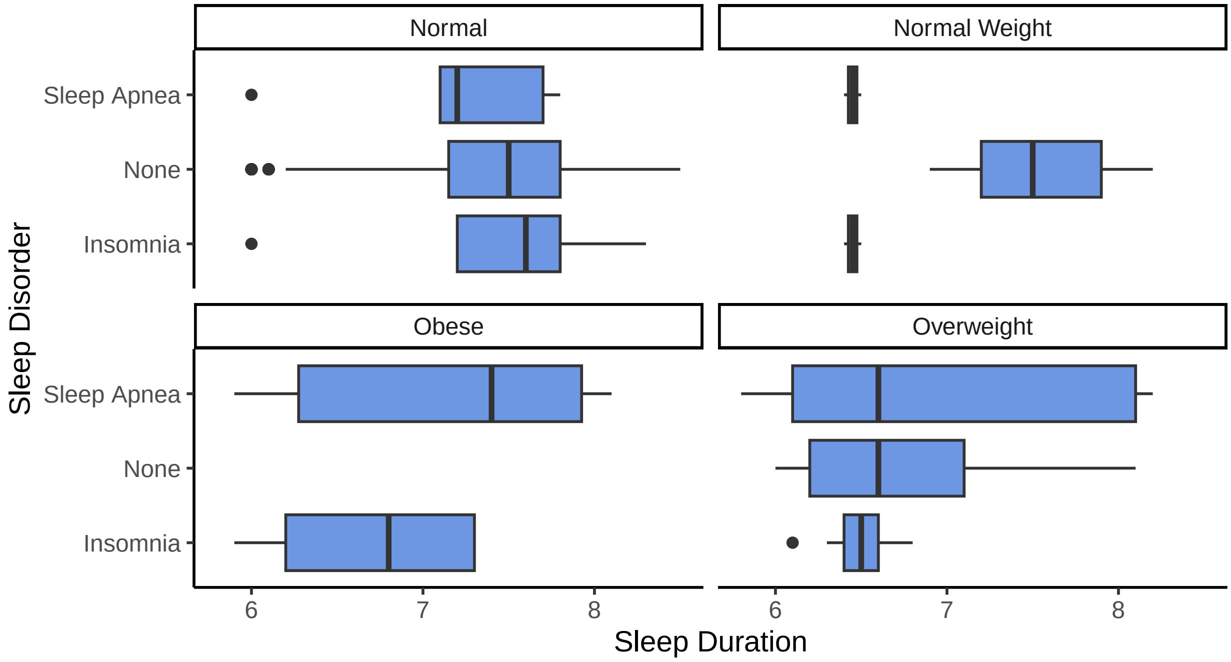
Figure 2:How Sleep Disorder Affect Stress Level



To begin with, the first graph demonstrates how sleep disorders affect people's stress levels. For example, people who are extremely stressed tend to have Sleep Apnea than Insomnia. However, although the total count also decreased, for just one level decrease in stress, people with Insomnia outnumbered those with Sleep Apnea. On the other hand, we can generalize and conclude that both seven and eight are considered extremely stressed, and the reason that people are not selecting stress levels nine and ten is response bias.

Another intriguing finding is that people who have sleep Apnea tend to be in two extremes and people with Insomnia are more evenly spread. This can be because people with Insomnia are growing stressed from time to time, and Sleep Apnea either significantly increases stress or does not affect the stress level at all. One thing to not forget is that only 30% of the extremely stressed population does not have sleep disorders, which is evidence proving that stress levels can lead to sleep disorders.

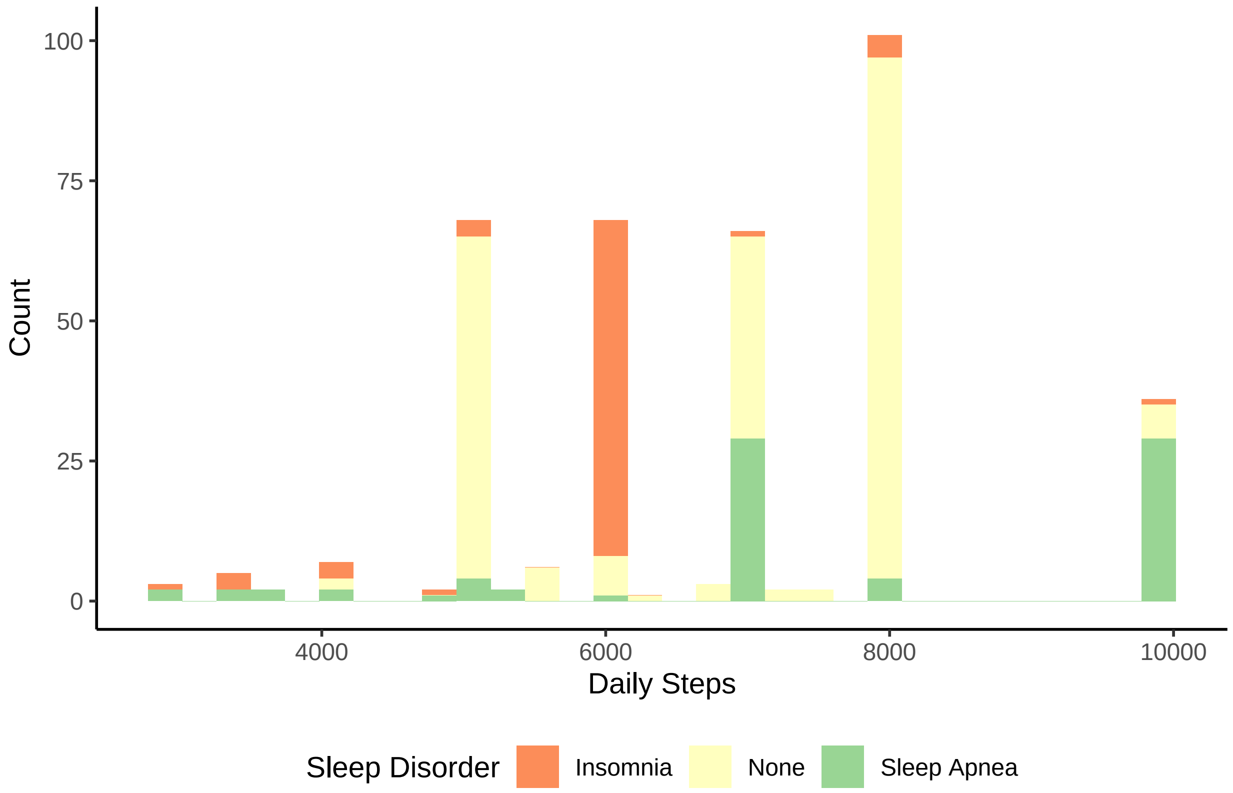
Figure 3: Sleep Duration Among Group With Different BMI Category and Sleep Disorder



The second graph presented a direct comparison of sleep duration between different groups based on sleep disorders and BMI category. From the graph, it's very easy to tell that the BMI category has four different classes with the "Normal" class being the largest (195), "Overweight" second (148), "Normal Weight" third (21), and "Obese" the fewest (10). Oddly, the author addresses the BMI category in four classes, but every piece of data is important in this analysis. On the other hand, It's clear that people who classify as obese either have sleep apnea or insomnia, which suggests that there can be a positive correlation between weight and sleep disorder; however, the reason can be because of the limited data.

In addition, normal people with insomnia had a higher median sleep duration than any other group, this suggests that people classified as normal weight can be less affected by insomnia. However, because the distribution of the data is not normal, and that means median is no longer a great indicator of the average. Moreover, people who classified as overweight and insomnia seem to have a relatively lower average of sleep duration compare to other people who also classified as insomnia.

Figure 4: Histogram of Daily Steps in Relation to Sleep Disorder

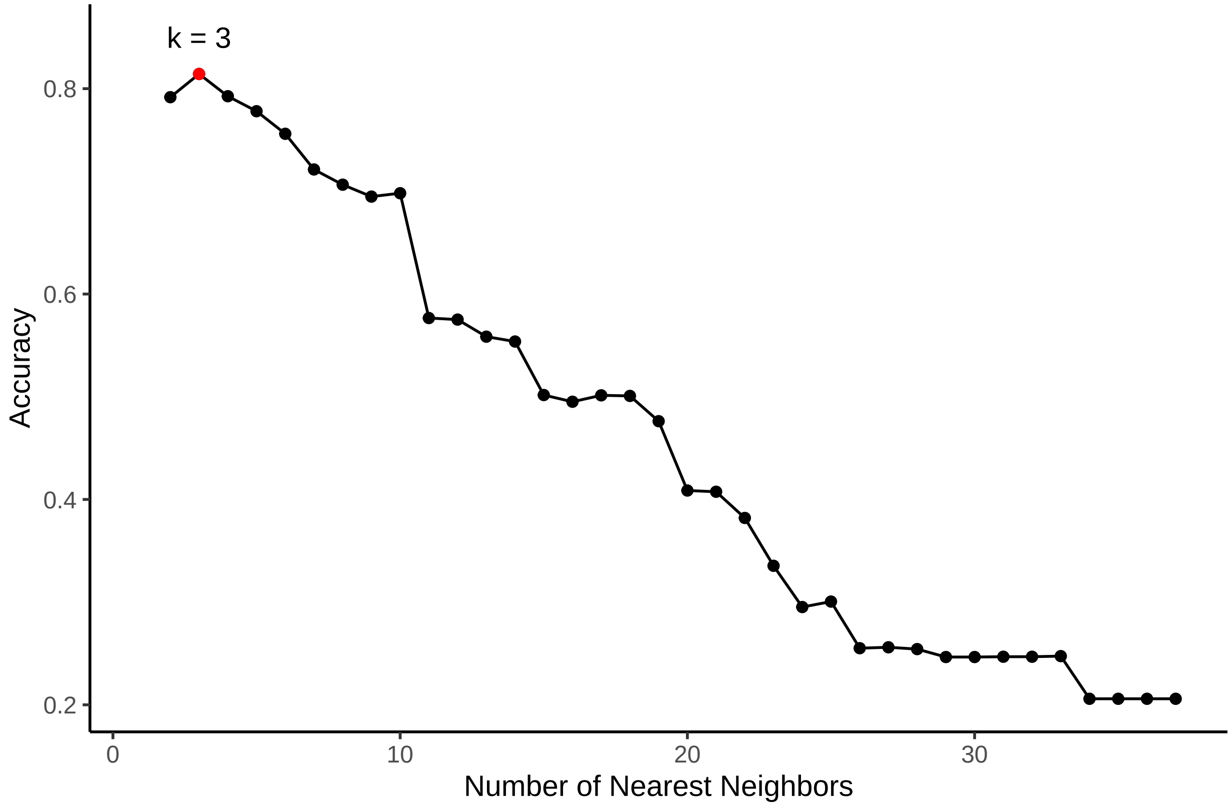


The third graph is a histogram of daily steps with sleep disorders as fills. One interesting finding about this graph everyone who walks less than 4,000 steps per day has some sort of sleep disorder, and people who walk the most seem to be dominated by sleep apnea with a small proportion of insomnia. This suggests that either lack of exercise has some kind of association and people with sleep apnea are not that severe that prevent daily activities.

KNN Classification

Next is to implement the actual KNN classification algorithm, the key to KNN classification is that it only support qualitative variables and require scaling to make sure every variable is compared on the same scale. This report will also run a 10-fold cross validation loop to examine the best k (Greatest accuracy).

Figure 5: Graph of the effect of the choice of k



Base on figure 5, the accuracy of the model decreases as k increases and the model performs best when k = 3.

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Insomnia None Sleep Apnea
##   Insomnia          15    2           2
##   None               2   42           1
##   Sleep Apnea        1    1           9
##
## Overall Statistics
##
##                Accuracy : 0.88
##                  95% CI : (0.7844, 0.9436)
##     No Information Rate : 0.6
##     P-Value [Acc > NIR] : 9.334e-08
##
##                   Kappa : 0.7841
##
##  Mcnemar’s Test P-Value : 0.9536
##
## Statistics by Class:
##
##                      Class: Insomnia Class: None Class: Sleep Apnea
## Sensitivity                   0.8333      0.9333             0.7500
## Specificity                   0.9298      0.9000             0.9683
## Pos Pred Value                0.7895      0.9333             0.8182
## Neg Pred Value                0.9464      0.9000             0.9531
## Prevalence                    0.2400      0.6000             0.1600
## Detection Rate                0.2000      0.5600             0.1200
## Detection Prevalence          0.2533      0.6000             0.1467
## Balanced Accuracy             0.8816      0.9167             0.8591
```

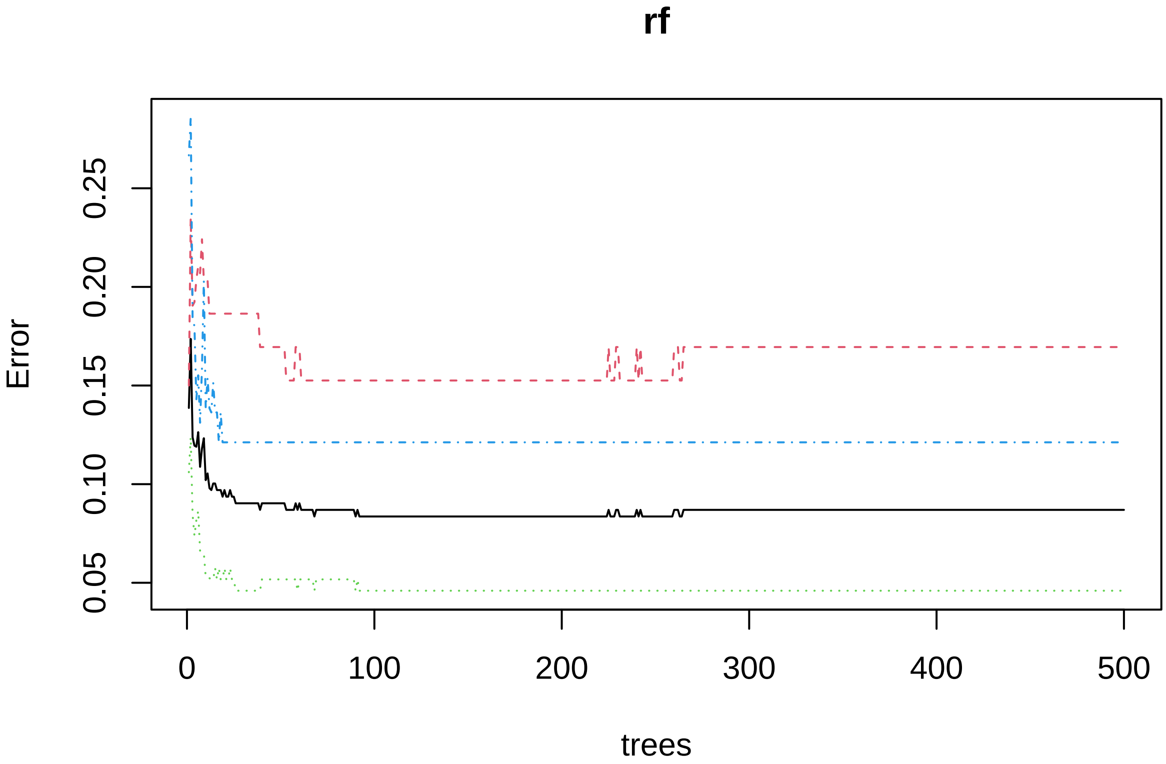
Above is a confusion matrix of the prediction of the final knn model using a random seed of 123 and a 8/2 test/train split. Based on the confusion matrix of the model, the accuracy is 88% with a 95% confidence interval of 78.44% to 94.36%. The confusion matrix also presents relevant statistics like sensitivity, specificity of the different classes of sleep disorder. Based on the statistics, the model is performing better in predicting "None" class than the two sleep disorders.

Random Forest

There's some advantages of using Random Forest when we are comparing against KNN classification. For example, random forest is an ensemble method that used out of bag (OOB) observations to calculate it's error rate, which means random forest models doesn't required cross-validation or train and test split because the error rate is automatically calculated through OOB observations. However, it's still important to create a train and test split using the same seed when we are comparing between different models.

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction    Insomnia None Sleep Apnea
##   Insomnia          15    0           1
##   None               2   44           1
##   Sleep Apnea        1    1          10
##
## Overall Statistics
##
##                Accuracy : 0.92
##                  95% CI : (0.834, 0.9701)
##     No Information Rate : 0.6
##     P-Value [Acc > NIR] : 4.645e-10
##
##                   Kappa : 0.8538
##
##  Mcnemar's Test P-Value : 0.5724
##
## Statistics by Class:
##
##                      Class: Insomnia Class: None Class: Sleep Apnea
## Sensitivity                   0.8333      0.9778             0.8333
## Specificity                   0.9825      0.9000             0.9683
## Pos Pred Value                0.9375      0.9362             0.8333
## Neg Pred Value                0.9492      0.9643             0.9683
## Prevalence                    0.2400      0.6000             0.1600
## Detection Rate                0.2000      0.5867             0.1333
## Detection Prevalence          0.2133      0.6267             0.1600
## Balanced Accuracy             0.9079      0.9389             0.9008
```

Above are the confusion matrix of the random forest model, which demonstrated to have a higher accuracy rate compared to the knn model by 4%. In addition, the kappa score of 0.85 also indicates more reliability for the random forest model as the kappa score evaluates the time something happened due to random chance, and higher kappa score means it's less likely because of a random chance.

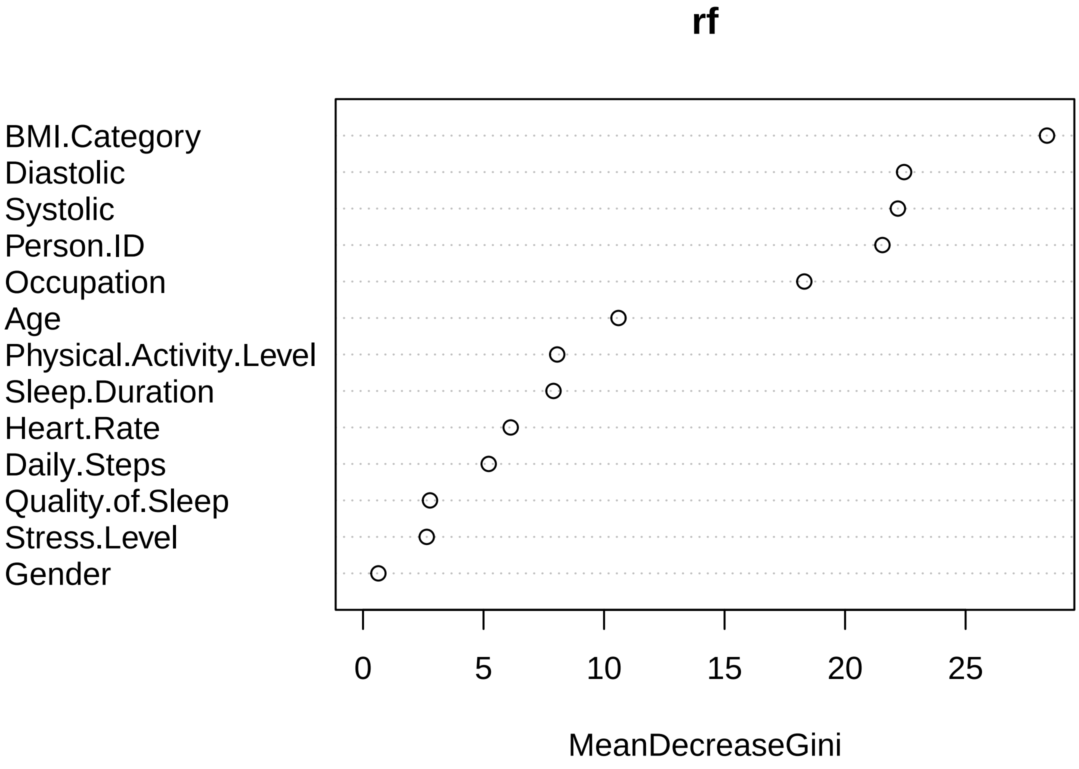


Above is a plot of the OOB errors produced by the model, the black line is the overall OOB error and the other three colored line represents each individual class. From this plot, we can tell that the OOB error for the model is very stable after 200th trees which suggest low variance of the model.

Table 2: Variable Importance Based on Gini Index

	MeanDecreaseGini
Person.ID	21.5498398
Gender	0.6367064
Age	10.5992922
Occupation	18.3084918
Sleep.Duration	7.9024390
Quality.of.Sleep	2.7783656
Physical.Activity.Level	8.0558331
Stress.Level	2.6429920
BMI.Category	28.3770816
Heart.Rate	6.1269512
Daily.Steps	5.2143580
Systolic	22.1907771
Diastolic	22.4479368

Table 2 provides the variable importance in the random forest model using Gini Index.



Above is a plot that sort the variables using the Gini index, and the plot demonstrated that BMI category is of the highest importance among all other variables.

Conclusion

This analysis discussed the different data visualizations using the "sleep" data set that contributed to existing findings and introduce new research questions, while compared the performance of KNN classification model and random forest model on classifying the sleep disorders based on the reminding variables. For example, the data visualizations in the EDA section describes a close connection between one's stress level and the sleep disorders, and a connection between the BMI category and the sleep disorder. In addition, the analysis also briefly talked about the unique aspects and details of using a KNN classification model and Random Forest model, and compare their performance using confusion matrix. From the results, it's clear that Random Forest model outperform KNN classification in many different ways. Included but not limited to, the overall accuracy, kappa score, and the time cost it takes to run both the algorithm.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
rm(list = ls()) # Clean environment
# Load libraries
library(psych)
library(tidyverse)
library(kableExtra)
library(modelr)
library(stats)
library(esquisse)
library(float)
library(caret)
library(FNN)
library(randomForest)
library(reprtree)

sleep <- read.csv("sleep.csv")

sleep <- sleep %>%
  separate(Blood.Pressure, c("place1", "place2")) %>% # Separate blood pressure into systolic and diast
  mutate(
    Systolic = as.integer(place1),
    Diastolic = as.integer(place2)
  ) %>%
  select(-place1, -place2)

quant_list <- c("Age", "Sleep.Duration", "Quality.of.Sleep", "Physical.Activity.Level", "Stress.Level",

summary_stat <- psych::describe(sleep[quant_list], skew = FALSE) %>%
  round(2)

summary_stat %>%
  kable(
    caption = 'Basic statistics of quantitative variables',
    booktabs = TRUE,
    align = c('l', rep('c', 8))
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c('striped', 'condensed'),
    font_size = 10,
    latex_options = "HOLD_position"
  )

ggplot(sleep) +
  aes(x = Stress.Level, fill = Sleep.Disorder) +
  geom_bar() +
  scale_fill_manual(
    values = c(Insomnia = "#F8766D",
    None = "#00C19F",
    `Sleep Apnea` = "#FF61C3")
  ) +
  labs(
    x = "Stress Level",
    y = "Count",
    title = "Figure 2:How Sleep Disorder Affect Stress Level",
    fill = "Sleep Disorder"
  ) +
  theme_classic() +
  theme(plot.title = element_text(size= 9)) +
  theme(plot.title = element_text(hjust = 0.5))

stress_population <- sleep %>%
  filter(
    Stress.Level >= 7
  )

ggplot(sleep) +
  aes(x = Sleep.Duration, y = Sleep.Disorder) +
  geom_boxplot(fill = "#6D97E2") +
  labs(
    title = "Figure 3: Sleep Duration Among Group With Different BMI Category and Sleep Disorder",
    x = "Sleep Duration",
    y = "Sleep Disorder"
  ) +
  theme_classic() +
  facet_wrap(vars(BMI.Category)) +
  theme(plot.title = element_text(size= 9)) +
  theme(plot.title = element_text(hjust = 0.5))

ggplot(sleep) +
  aes(x = Daily.Steps, fill = Sleep.Disorder) +
  geom_histogram(bins = 30L) +
  scale_fill_brewer(palette = "Spectral", direction = 1) +
  labs(
    x = "Daily Steps",
    y = "Count",
    title = "Figure 4: Histogram of Daily Steps in Relation to Sleep Disorder",
    fill = "Sleep Disorder"
  ) +
  theme_classic() +
  theme(legend.position = "bottom") +
  theme(plot.title = element_text(size= 9)) +
  theme(plot.title = element_text(hjust = 0.5))

# Scaled the data
xvars <- names(sleep)
xvars <- xvars[-c(1, 12)]

# Setting up dummy variables (KNN Only works with quantitative variables)
sleep_scaled <- sleep %>%
  mutate(
    Gender = ifelse(Gender == "Male", 1, 0),

    Occupation = case_when(
      Occupation == "Software Engineer" ~ 0,
      Occupation == "Doctor" ~ 1,
      Occupation == "Sales Representative" ~ 2,
      Occupation == "Teacher" ~ 3,
      Occupation == "Nurse" ~ 4,
      Occupation == "Engineer" ~ 5,
      Occupation == "Accountant" ~ 6,
      Occupation == "Scientist" ~ 7,
      Occupation == "Lawyer" ~ 8,
      Occupation == "Salesperson" ~ 9,
      Occupation == "Manager" ~ 10),

    BMI.Category = case_when(
      BMI.Category == "Overweight" ~ 0,
      BMI.Category == "Normal" ~ 1,
      BMI.Category == "Obese" ~ 2,
      BMI.Category == "Normal Weight" ~ 3,
    )
  )

sleep_scaled[ , xvars] <- scale(sleep_scaled[ , xvars],
                            center = TRUE,
                            scale = TRUE)

# Assign fold value and rearrange fold
num_folds <- 10
folds <- cut(x = 1:nrow(sleep_scaled), breaks = num_folds, labels = FALSE)

set.seed(123)
folds <- sample(folds)
set.seed(NULL)

# Create a matrix to store validation output
maxK <- 37 # Only 37 patterns
accuracy_mat <- matrix(NA, nrow = num_folds, ncol = maxK)

# Perform validation
for(i in 1:num_folds){
  train_ind <- which(folds == i)
  train <- sleep_scaled[train_ind, ]
  test <- sleep_scaled[-train_ind, ]
  for(j in 2:maxK){
    knn_model <- knn(train = train[ , xvars, drop = FALSE],
               test = test[ , xvars, drop = FALSE],
               cl = train$Sleep.Disorder,
               k = j)

    test <- test %>%
          mutate(pred_sleep_disorder = knn_model)

    confusion.mat = table(test$pred_sleep_disorder, test$Sleep.Disorder)

    accuracy <- sum(diag(confusion.mat))/sum(confusion.mat)

    accuracy_mat[i, j] <- accuracy
  }
}

accuracy_vec <- colMeans(accuracy_mat) # Convert matrix to vector by finding the average of each fold

temp_df <- data.frame(NN = 1:37, accuracy = accuracy_vec)

ggplot(data = temp_df, mapping = aes(x = NN, y = accuracy)) +
  geom_point() +
  geom_line() +
  labs(title = "Figure 5: Graph of the effect of the choice of k",
       x = "Number of Nearest Neighbors",
       y = "Accuracy") +
  theme_classic() +
  theme(plot.title = element_text(size= 9)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point(data = temp_df %>% filter(accuracy == max(accuracy[2:37])),
      pch = 21, colour = "red", fill = "red") +
  annotate("text", x = 3, y = 0.85, label = "k = 3")

set.seed(123)
train_ind <- sample(1:nrow(sleep_scaled), floor(0.8*nrow(sleep_scaled)))
set.seed(NULL)

train <- sleep_scaled[train_ind, ]
test <- sleep_scaled[-train_ind, ]

final_knn <- knn_model <- knn(train = train[ , xvars, drop = FALSE],
               test = test[ , xvars, drop = FALSE],
               cl = train$Sleep.Disorder,
               k = 3)

test <- test %>%
          mutate(pred_sleep_disorder = knn_model)

knn_confusion.mat = caret::confusionMatrix(data = test$pred_sleep_disorder,
                                           reference = as.factor(test$Sleep.Disorder))
knn_confusion.mat
```

```r
# New testing and training set
train_tree <- sleep[train_ind, ]
test_tree <- sleep[-train_ind, ]

# Build a random forest to predict sleep disorder, default of ntree = 500 and mtry = round up sqrt(# of
rf <- randomForest(as.factor(Sleep.Disorder) ~ ., data = train_tree, ntree = 500, mtry = 4)

# Create a vector of prediction probability
pred_rf <- predict(rf, newdata = test_tree, type = "prob")

# Create a confusion matrix
pred_vec <- predict(rf, newdata = test_tree, type = "response")
rf_confusion.mat = caret::confusionMatrix(data = as.factor(pred_vec),
                                          reference = as.factor(test$Sleep.Disorder))
rf_confusion.mat

plot(rf)

importance(rf) %>%
    kable(
    caption = 'Variable Importance Based on Gini Index',
    booktabs = TRUE,
    align = c('l', rep('c', 2))
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c('striped', 'condensed'),
    font_size = 10,
    latex_options = "HOLD_position"
  )

varImpPlot(rf)
```