Quality of Sleep: A Study on How You Could Sleep Better

Marijan Sorić Marco Pozzetto Solène Couchot

15 April, 2024

Abstract

Is sleeping more really means better quality of sleep? The purpose of the project is to predict the quality of sleep for people based on health data collected. We tackle this aim through supervised learning using statistical learning techniques and methods. Beyond the goal of predicting a target Y based on data X, we also want to analyse the influence of covariates (features) on the response Y so that we can conclude on advise over people based on who they are. The data set contains 374 examples with 13 columns, including the response one. Among these covarites, there are different types of information about each person. There are quantitative and qualitative (categorical) covariates. We used backward selection of covariates in features selection. To perform the regression task (quality of sleep prediction), we developed different models and compare their performance on a test set with the Mean Absolute Error metric. At first, we build very simple and naive models called 'baseline model' to have an idea about which performance can we expect, and how well more sophisticated models performs. We use a linear regression and a tree based models (including Random Forest). We can rank important variables as following: sleep duration, stress level, age, daily steps. As expected, the sleep duration is one of the most important features, but other are significant and should be considered. It is surprising easy to obtain a model that performs well with the MAE metric on the test set. However, this project isn't only about prediction but also inference.

Introduction

The idea of our problem is to have a better understanding of on which features (external or internal factors) influence, and how much, our quality of sleep. To do so, we based our study on the data set "Sleep Health and Lifestyle Dataset" by the user Laksika Tharmalingam on Kaggle. The response variable Y used is `Quality.of.Sleep`. It is a subjective rating (integer values) of the quality of sleep, ranging from 1 to 10 ($Y \in \{1, .., 10\}$). The goals of the project is the following:

- Prediction task: build a model that predict the quality of sleep of a new person $\widehat{Y}$, given some data.
- Inference task: understand which features are the most important, and on which ones we can do something. *By looking at coefficients for example.*

At first, we decided to tackle this problem as a regression and not a classification, as there is a relationship (order) between classes. As a matter of fact, we want to take into account how far was the prediction from the "ground truth". If we would use misclassification rate, we loose information about how precise is our prediction. However, as we use regressor, the output is a real number $\widehat{Y}_R \in [1, 10]$. Finally, as a test metric, we decided to use Mean Average Error[1], and considered the closest integer as model's output: $\widehat{Y} := \lfloor \widehat{Y}_R \rceil \in \{1, .., 10\}$. We also notice that rounding up the output of the model improve the MAE score[2] when it's performing well enough.

In the tree based, we use classification task because it was easier to implement. Moreover, when the algorithm has to decide what is the output for a sub sample of data point, a majority vote for classification seems to be

[1] Mean Average Error: $\frac{1}{N_{\text{test}}} \sum |\hat{y}_i - y_i|$

[2] This is an intuitive result. Let take a regression that performs well. Let $\widehat{Y} \sim \mathcal{U}(]0.6, 1.4[)$ for the the target $Y = 1$, then $\left\lfloor \widehat{Y} \right\rceil = 1$. If we compute the L^1 loss, we have: $\mathbb{E}_{\widehat{Y}} \left| \lfloor \widehat{Y} \rceil - Y \right| \leq \mathbb{E}_{\widehat{Y}} \left| \widehat{Y} - Y \right|$. As a matter of fact, $\mathbb{E}_{\widehat{Y}} \left| \lfloor \widehat{Y} \rceil - Y \right| = 0$, while $\mathbb{E} \left| \widehat{Y} - Y \right| = \int_{0.6}^{1.4} |\hat{y} - 1| \, f_{\widehat{Y}}(\hat{y}) \mathrm{d}\hat{y} = 2 \int_0^{0.4} x \frac{1}{1.4 - 0.6} \mathrm{d}x > 0$

similar to an average over these sample because we decide to round up the output. This prediction problem had been tackled trough different angles, but it might also be seen as an ordered classification...

Descriptive data analysis/statistics

Data description

The data contains 374 examples with 13 columns, including the response one Among these covarites, there are different types of information:

- Who is the person: gender, age.
- What does the person do: occupation (work), physical activity level, daily steps
- Medical data: heart rate, blood pressure, BMI category, stress level
- Sleep related data: sleep duration, sleep disorder
- Response: `Quality.of.Sleep`

It will be interesting to understand which "types" of data are the most important in inference.

Data preprocessing

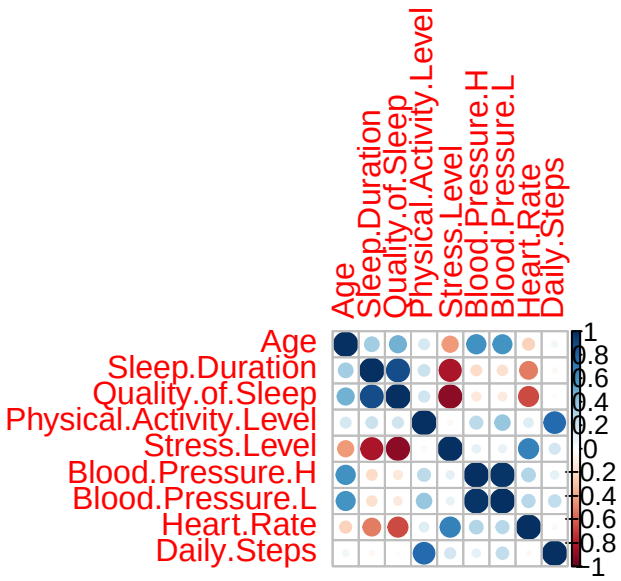
Firstly, we have modified our dataset by splitting the blood pressure information into two separate variables: low blood pressure level and high blood pressure level. Additionally, we have adjusted the BMI category column: specifically, we have changed the values "normal Weight" to simply "normal" as the distinction was not specified, thereby retaining the categories of "normal", "overweight", and "obese". Lastly, we have removed the ID column as it is not relevant for our models. We decided not to scale the data. As a matter of fact, it is not necessary with linear regression, and useless for tree models.

Data visualization

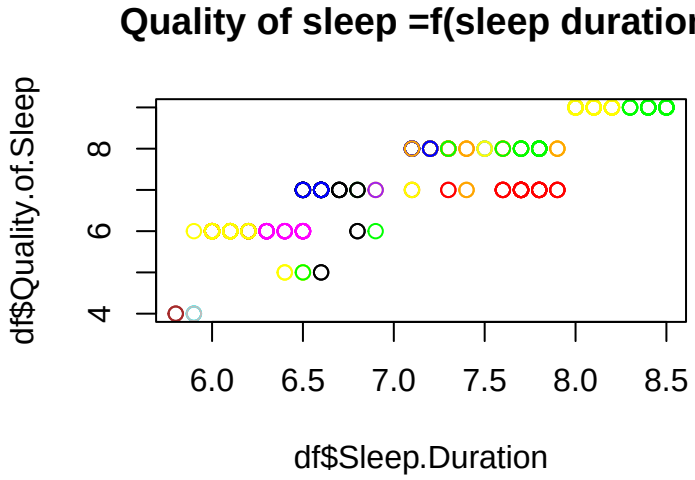
Now we can make an overview of the dataset:

```
##    Person.ID         Gender               Age         Occupation
##  Min.   :  1.00   Length:374         Min.   :27.00   Length:374
##  1st Qu.: 94.25   Class :character   1st Qu.:35.25   Class :character
##  Median :187.50   Mode  :character   Median :43.00   Mode  :character
##  Mean   :187.50                      Mean   :42.18
##  3rd Qu.:280.75                      3rd Qu.:50.00
##  Max.   :374.00                      Max.   :59.00
##  Sleep.Duration  Quality.of.Sleep Physical.Activity.Level  Stress.Level
##  Min.   :5.800   Min.   :4.000    Min.   :30.00           Min.   :3.000
##  1st Qu.:6.400   1st Qu.:6.000    1st Qu.:45.00           1st Qu.:4.000
##  Median :7.200   Median :7.000    Median :60.00           Median :5.000
##  Mean   :7.132   Mean   :7.313    Mean   :59.17           Mean   :5.385
##  3rd Qu.:7.800   3rd Qu.:8.000    3rd Qu.:75.00           3rd Qu.:7.000
##  Max.   :8.500   Max.   :9.000    Max.   :90.00           Max.   :8.000
##  BMI.Category       Blood.Pressure.H Blood.Pressure.L   Heart.Rate
##  Length:374         Min.   :115.0    Min.   :75.00    Min.   :65.00
##  Class :character   1st Qu.:125.0    1st Qu.:80.00    1st Qu.:68.00
##  Mode  :character   Median :130.0    Median :85.00    Median :70.00
##                     Mean   :128.6    Mean   :84.65    Mean   :70.17
##                     3rd Qu.:135.0    3rd Qu.:90.00    3rd Qu.:72.00
##                     Max.   :142.0    Max.   :95.00    Max.   :86.00
##   Daily.Steps    Sleep.Disorder
##  Min.   : 3000   Length:374
##  1st Qu.: 5600   Class :character
```

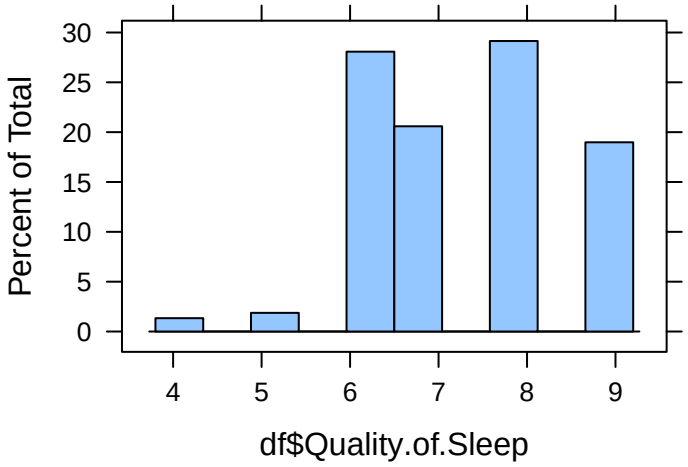
```
##  Median : 7000   Mode  :character
##  Mean   : 6817
##  3rd Qu.: 8000
##  Max.   :10000
```



The matrix of correlation between features that are non-categorical, shed light on the correlation between the response and mainly the `Quality.of.Sleep`, `Strees.Level` and `Heart.Rate`. This trend will be captured later with the linear regressor. One can notice that some features are correlated with each other: `Blood.Pressure.H` and `Blood.Pressure.L`, that is why we might keep only one in order to decrease the dimension.



This plot underlines the correlation between the quality of sleep and the duration of sleep. We can draw a trend but there exists a range of quality of sleep for some sleep duration. In colors, we represented the 11 occupations (jobs). We can notice that they are not that much about the quality of sleep. Plus, some classes are very poorly represented. That is why we might remove this covariate soon.



The response variable is an integer, mostly between 6 and 9. We could expect a gaussian distribution but it is not the case. It may be because of the subjectivity of this feature.

Data split

We randomly split the data set into two parts: a training set (80%) and a test set (20%), in order to be able to compare our different models on the prediction task. On the training set, each method will use differently the training set, by splitting again to obtain a validation set for hyperparameters tuning for example.

Methods

To asset the prediction problem we implemented different models and compare them with the MAE metric. That way, we take in account the precision of the prediction, as discuss previously.

Baseline (naive)

Before applying any method from statistical learning, we decided to build a very simple and naive model as a baseline. This baseline will give an "MAE to beat". That is why we develop multiple very simple models.

- The first one is given by the formula: $\widehat{Y}_{BL} = \left\lfloor \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} Y_i \right\rceil$. The output is always the same: the mean over all the response Y_i of the training set.
- The second baseline (better approximation), is considering the mean over the subset $C(occ)$ of people that have the same *occupation* (variable `Occupation`): $\widehat{Y}_{BL}(occ) = \left\lfloor \frac{1}{|C(occ)|} \sum_{i \in C(occ)} Y_i \right\rceil$

Finally, we tried another small model, very simple and that is linked with the next part: a linear regression with only one covariate: `Sleep.Duration`.

Linear Regression

Now we want to analyze the quality of sleep using a multiple linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$ where X_j is the jth predictor and β_j the respective regression coefficient. First of all, we consider the full model:

```
##
## Call:
## lm(formula = trainLabels ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.99983 -0.10038 -0.00248  0.07678  0.84807
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.052e+01  1.138e+00   9.248  < 2e-16 ***
## GenderMale                      7.044e-01  8.220e-02   8.569 7.42e-16 ***
## Age                             7.312e-02  5.646e-03  12.951  < 2e-16 ***
## OccupationDoctor               -4.577e-01  9.693e-02  -4.722 3.72e-06 ***
## OccupationEngineer             -6.116e-01  8.837e-02  -6.921 3.12e-11 ***
## OccupationLawyer               -3.107e-01  1.116e-01  -2.785 0.005726 **
## OccupationManager              -2.939e-01  2.326e-01  -1.264 0.207389
## OccupationNurse                -1.575e-01  1.093e-01  -1.441 0.150578
## OccupationSales Representative -1.040e+00  2.465e-01  -4.219 3.33e-05 ***
## OccupationSalesperson          -9.359e-01  9.834e-02  -9.517  < 2e-16 ***
## OccupationScientist            -3.374e-01  1.625e-01  -2.077 0.038750 *
## OccupationSoftware Engineer    -4.214e-01  1.628e-01  -2.589 0.010129 *
## OccupationTeacher              -4.699e-01  8.029e-02  -5.853 1.37e-08 ***
## Sleep.Duration                  1.512e-01  5.163e-02   2.929 0.003679 **
## Physical.Activity.Level        -9.025e-05  1.587e-03  -0.057 0.954690
## Stress.Level                   -3.684e-01  2.898e-02 -12.710  < 2e-16 ***
## BMI.CategoryObese              -1.308e-01  2.443e-01  -0.535 0.592891
## BMI.CategoryOverweight         -4.645e-01  9.878e-02  -4.703 4.05e-06 ***
## Blood.Pressure.H               -2.202e-02  1.690e-02  -1.303 0.193591
## Blood.Pressure.L               -7.349e-04  2.212e-02  -0.033 0.973519
## Heart.Rate                     -3.695e-02  1.008e-02  -3.666 0.000295 ***
## Daily.Steps                     2.873e-05  2.182e-05   1.317 0.188984
## Sleep.DisorderNone              1.731e-01  5.868e-02   2.950 0.003445 **
## Sleep.DisorderSleep Apnea       1.557e-01  6.559e-02   2.374 0.018273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2133 on 277 degrees of freedom
## Multiple R-squared:  0.9716, Adjusted R-squared:  0.9692
## F-statistic: 411.9 on 23 and 277 DF,  p-value: < 2.2e-16
```

From the F-statistic, it's evident that at least one of the predictors contributes significantly to the model, and the high R^2 indicates that the model adequately explains the response variable.

However, upon examining the p-values, we observe that some covariates lack significance. To refine the model and reduce complexity, we can employ a backward stepwise selection approach, iteratively removing the least useful predictor.

In the initial step, we opt to remove the low blood pressure level variable, as it exhibits high correlation with the high blood pressure level.

```
g1=update(g,.~. -Blood.Pressure.L)
summary(g1)
```

```
##
## Call:
## lm(formula = trainLabels ~ Gender + Age + Occupation + Sleep.Duration +
##     Physical.Activity.Level + Stress.Level + BMI.Category + Blood.Pressure.H +
##     Heart.Rate + Daily.Steps + Sleep.Disorder, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.99931 -0.10056 -0.00255  0.07704  0.84800
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.053e+01  1.111e+00   9.482  < 2e-16 ***
## GenderMale                      7.048e-01  8.106e-02   8.694 3.09e-16 ***
## Age                             7.321e-02  5.071e-03  14.435  < 2e-16 ***
## OccupationDoctor               -4.581e-01  9.593e-02  -4.775 2.91e-06 ***
## OccupationEngineer             -6.114e-01  8.794e-02  -6.952 2.56e-11 ***
## OccupationLawyer               -3.104e-01  1.111e-01  -2.794 0.005567 **
## OccupationManager              -2.935e-01  2.318e-01  -1.266 0.206563
## OccupationNurse                -1.584e-01  1.061e-01  -1.493 0.136676
## OccupationSales Representative -1.040e+00  2.460e-01  -4.227 3.21e-05 ***
## OccupationSalesperson          -9.356e-01  9.786e-02  -9.561  < 2e-16 ***
## OccupationScientist            -3.370e-01  1.618e-01  -2.083 0.038136 *
## OccupationSoftware Engineer    -4.220e-01  1.615e-01  -2.613 0.009459 **
## OccupationTeacher              -4.697e-01  7.999e-02  -5.872 1.23e-08 ***
## Sleep.Duration                  1.508e-01  4.949e-02   3.046 0.002542 **
## Physical.Activity.Level        -9.229e-05  1.583e-03  -0.058 0.953550
## Stress.Level                   -3.682e-01  2.844e-02 -12.948  < 2e-16 ***
## BMI.CategoryObese              -1.306e-01  2.438e-01  -0.536 0.592572
## BMI.CategoryOverweight         -4.662e-01  8.531e-02  -5.465 1.03e-07 ***
## Blood.Pressure.H               -2.255e-02  6.135e-03  -3.675 0.000286 ***
## Heart.Rate                     -3.697e-02  1.005e-02  -3.677 0.000283 ***
## Daily.Steps                     2.838e-05  1.908e-05   1.488 0.138014
## Sleep.DisorderNone              1.734e-01  5.816e-02   2.981 0.003127 **
## Sleep.DisorderSleep Apnea       1.558e-01  6.541e-02   2.382 0.017869 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2129 on 278 degrees of freedom
## Multiple R-squared:  0.9716, Adjusted R-squared:  0.9693
## F-statistic: 432.2 on 22 and 278 DF,  p-value: < 2.2e-16
```

As anticipated, R^2 decreases as we remove variables, reflecting the reduction in explanatory power with fewer predictors. However, R^2_{adj} increases, consistent with its role in penalizing the addition of variables that do not contribute significantly to model improvement.

For similar reasons, we can consider removing the variable representing physical activity level, given its high correlation with daily steps.

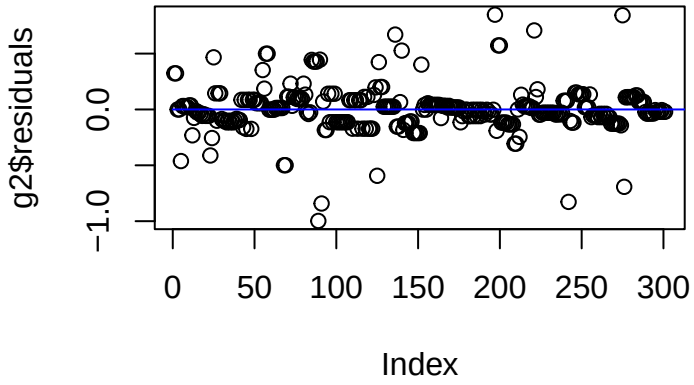
```
g2=update(g1,.~. -Physical.Activity.Level)
summary(g2)
```

```
##
## Call:
## lm(formula = trainLabels ~ Gender + Age + Occupation + Sleep.Duration +
##     Stress.Level + BMI.Category + Blood.Pressure.H + Heart.Rate +
##     Daily.Steps + Sleep.Disorder, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99786 -0.10082 -0.00265  0.07760  0.84832
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

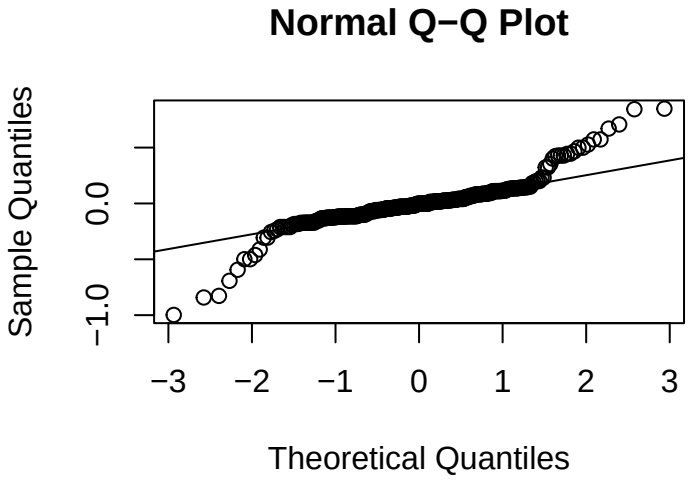
```
## (Intercept)                      1.056e+01  1.032e+00  10.227  < 2e-16 ***
## GenderMale                       7.031e-01  7.575e-02   9.282  < 2e-16 ***
## Age                              7.322e-02  5.060e-03  14.470  < 2e-16 ***
## OccupationDoctor                -4.560e-01  8.864e-02  -5.144 5.06e-07 ***
## OccupationEngineer              -6.096e-01  8.266e-02  -7.375 1.88e-12 ***
## OccupationLawyer                -3.081e-01  1.036e-01  -2.975 0.003187 **
## OccupationManager               -2.938e-01  2.314e-01  -1.270 0.205161
## OccupationNurse                 -1.584e-01  1.059e-01  -1.495 0.135948
## OccupationSales Representative -1.037e+00  2.391e-01  -4.335 2.04e-05 ***
## OccupationSalesperson           -9.339e-01  9.324e-02 -10.016  < 2e-16 ***
## OccupationScientist             -3.366e-01  1.614e-01  -2.086 0.037858 *
## OccupationSoftware Engineer     -4.187e-01  1.509e-01  -2.774 0.005915 **
## OccupationTeacher               -4.694e-01  7.967e-02  -5.892 1.10e-08 ***
## Sleep.Duration                   1.498e-01  4.663e-02   3.213 0.001469 **
## Stress.Level                    -3.680e-01  2.821e-02 -13.046  < 2e-16 ***
## BMI.CategoryObese               -1.314e-01  2.430e-01  -0.541 0.588947
## BMI.CategoryOverweight          -4.658e-01  8.491e-02  -5.486 9.23e-08 ***
## Blood.Pressure.H                -2.261e-02  6.041e-03  -3.742 0.000221 ***
## Heart.Rate                      -3.711e-02  9.712e-03  -3.821 0.000164 ***
## Daily.Steps                      2.746e-05  1.069e-05   2.569 0.010718 *
## Sleep.DisorderNone               1.733e-01  5.805e-02   2.986 0.003079 **
## Sleep.DisorderSleep Apnea        1.561e-01  6.515e-02   2.396 0.017256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2125 on 279 degrees of freedom
## Multiple R-squared:  0.9716, Adjusted R-squared:  0.9695
## F-statistic: 454.4 on 21 and 279 DF,  p-value: < 2.2e-16
```

In this final model, we observe an increase in the R^2_{adj}, indicating an improvement in model fit while simultaneously reducing model complexity. Furthermore, all covariates appear to be significant in explaining the model.

The assumptions of the linear regression model are that the error terms are independent and identically distributed with a normal distribution, that is, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Thus, we can check the assumption by looking at the residuals:



From this plot we can say that the assumption of zero mean of the residuals is verified because the points are centered around zero. Moreover, the assumption of independence of residuals looks verified because there are no particular trend in the "cloud" of points.




```
##
##  Shapiro-Wilk normality test
##
## data:  g2$residuals
## W = 0.85269, p-value = 2.795e-16
```

From the QQ-diagram we can say that the assumption of normal distribution of the residuals is not verified because there is no a straight line. It can also be confirmed by computing the Shapiro test, which yields a very small p-value, leading us to reject the null hypothesis regarding the normal distribution of the residuals.

Since the model has no Gaussian residuals, we could consider transforming the data to make the residuals more Gaussian(e.g. Box-Cox transformation), or we could explore using non-parametric models like decision trees. Decision trees offer robustness to non-Gaussian residuals as they don't rely on specific distributional assumptions. They're flexible and interpretable, making them a viable alternative.

Tree based model

As a non-parametric method, we use trees and random forests. The idea of trees is to divide the space of covariates into a number of regions, and to each region assign a predictive value. We are computing classification trees, consequently in each region we use a majority vote to derive the predictive value. Our regions are derived step by step : at each step we divide one of the existing region into two regions with a criteria that depends on a single covariate (we consider whether the covariate greater than a certain deciding value). Which region is divided and what our deciding value is, is decided by minimizing a chosen distance (our splitting criterion) between the points in that region and its predictive value. We will consider (and choose between) the Gini index and cross entropy (we actually use a scaled version of the cross entropy) as splitting criterions. We must also decide on a stopping criterion for our algorithm : it can be the minimum number of points in a region or a minimum "gain" in precision. We can compute e very bushy tree and use cross validation prune it. In our tree, our nodes correspond to splitting criteria and branches to regions. Trees are easy to understand, plot and interpret. The tree automatically selects variables. However, large trees have a high variability; what's more they are easily influenced by change of data (low robustness). Consequently we also investigate bagging (a majority vote over predictions by over multiple trees; it decreases variability and increases robustness) and random forests (similar to bagging, but retristing the possible variable at each node, consequently decreasing variance and increasing randomnes). Those methods use multiple trees, consequently we can investigate variable importance. To evaluate the performance of our models and compare them to our other models, we use the MAE.

Results and interpretation

Baseline

Naive models get MAE score that seems already low. In average, the baseline model predicts one scale away from the target.

In fact, as almost all the response are around 7 and 8, according to the previous response histogram. We can notice that there are approximately 80 examples in the test set, so the lowest MAE that is not null is $\approx \frac{1}{80} = 0.0125$. Even for the baseline, the absolute error $\left|\widehat{Y}_{BL} - Y\right|$ is 0 or 1, but not more than that. Hence the mean square error seems useless.

```
##                            Model       MAE
## 1             Baseline Average 0.9863014
## 2 Baseline Occupation Average 0.5616438
```

Linear Regression

From the analysis of the final model obtained with the multiple linear regression, several noteworthy observations emerge:

- Individuals experiencing insomnia or other sleep-related disturbances tend to exhibit lower sleep quality.
- Obese or overweight individuals tend to have lower sleep quality.
- Among various occupations, sales representatives appear to have the most adverse impact on sleep quality.
- Additionally, the male gender shows a positive coefficient, suggesting a seemingly higher quality of life for males. However, in reality, the average sleep quality among males is lower than that among females. This apparent discrepancy may occur due to the summary of the linear model presenting a logically opposite result. Upon considering other covariates in the model, especially those potentially correlated with gender, the observed effect of gender may differ from its net effect when assessed individually. These findings underscore the importance of thorough consideration and interpretation of the model's results, particularly in contexts where covariates may interact or confound each other's effects.

Now we can compute the metric MAE by making prediction with the same test set used in the baseline method:

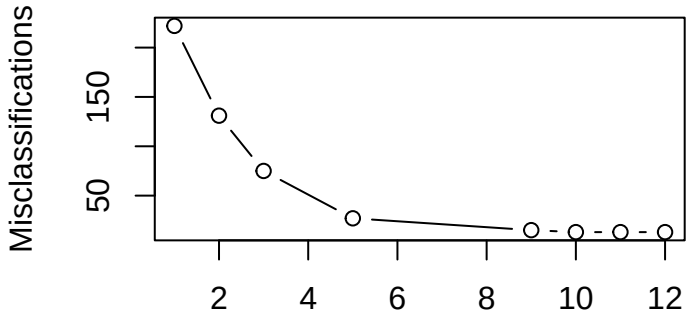
```
##                            Model        MAE
## 1             Baseline Average 0.98630137
## 2 Baseline Occupation Average 0.56164384
## 3           Linear Regression 0.09589041
```

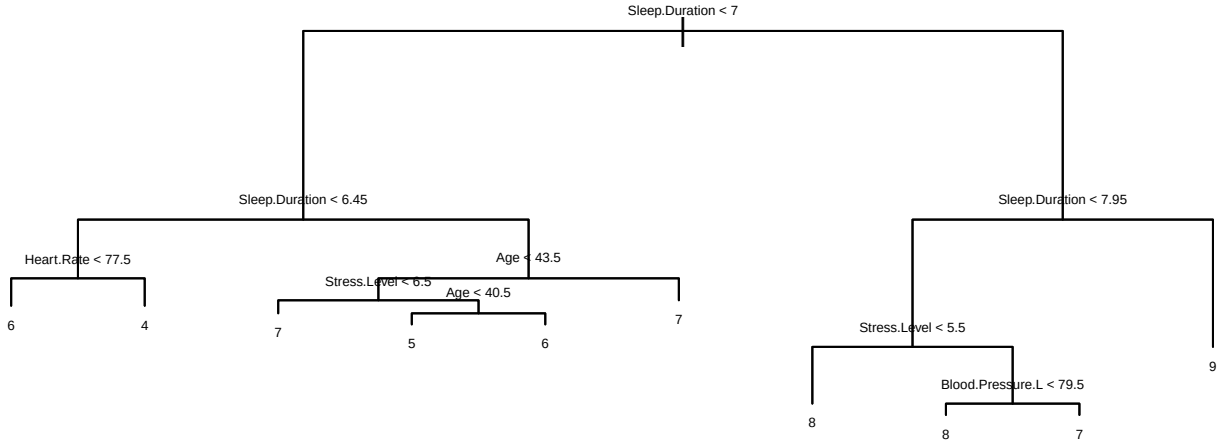
It means that the multiple linear regression has a much better prediction accuracy than the baseline method.

Tree based model

We first compute a single tree using our training data and all variables, both for cross entropy and Gini index. Among the variables appearing on the cross entropy tree, we have 'occupation doctor', which seem like a difficult factor to interpret. Consequently, we derive the same trees removing the categorical variables which seem most difficult to interpret ('occupation', 'BMI category' and 'sleep disorder'). In both cases, the tree computed using the deviance has a significantly lower MAE than the one computed using the Gini index (0.096 vs 0.397, and 0.096 vs 0.024 respectively). It seems that removing those variables doesn't increase error - in fact it diminishes it. While we are looking with less data, the structure of trees (construction step by step) means that such scenarios are not impossible. We now work without the variables 'occupation', 'BMI category' and 'sleep disorder'. Thus far we have worked with the default stopping criterion of the function 'tree'. We now define a stopping criterion allowing for a deeper tree. Using cross validation and with the misclassification rate as a reference (not the MAE), we compute how many terminal nodes of this new tree

we should keep. The graph of misclassifications by number of terminal nodes has us choosing a terminal number of nodes at 10. Coincidentally, this is the same tree we obtained with our previous parameters.






```
## [1] 0.08219178
```

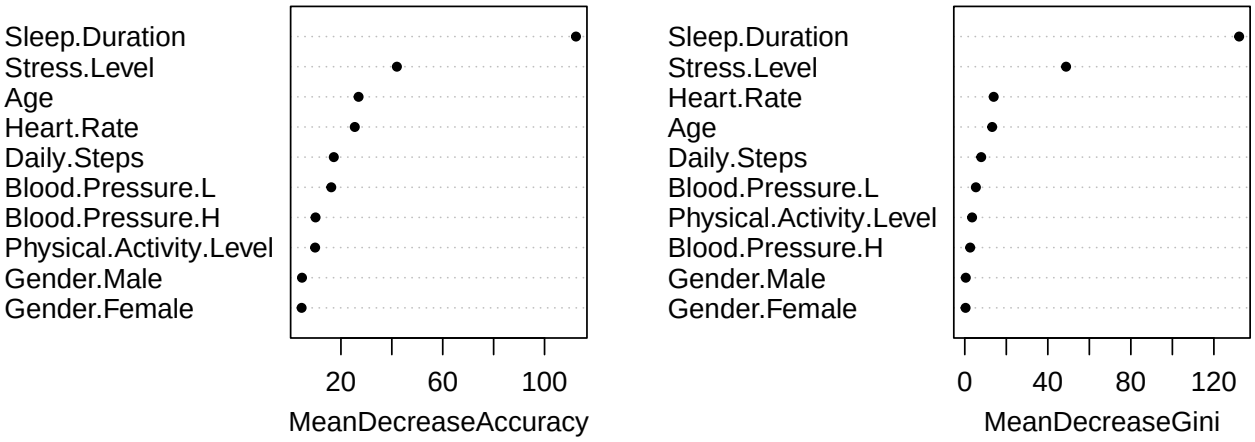
We see from the graph that the subtree with 5 nodes gives little more misclassifications than our 10 nodes tree; we could also have chosen to consider it. It seems clear from the tree we obtain that the duration of sleep is the main factor when it comes to estimating the quality of sleep (it is used for the first three nodes). Age, stress level, heart rate and blood pressure are also considered.

We now use bagging. As expected, this model gives us a lower MAE.

```
##
## yhat.bag  4  5  6  7  8  9
##        4  0  0  0  0  0  0
##        5  0  0  1  0  0  0
##        6  0  0 21  0  0  0
##        7  0  0  0 13  0  0
##        8  0  0  1  2 21  0
##        9  0  0  0  0  0 14
```

```
## [1] 0.06849315
```

r.QS.bag



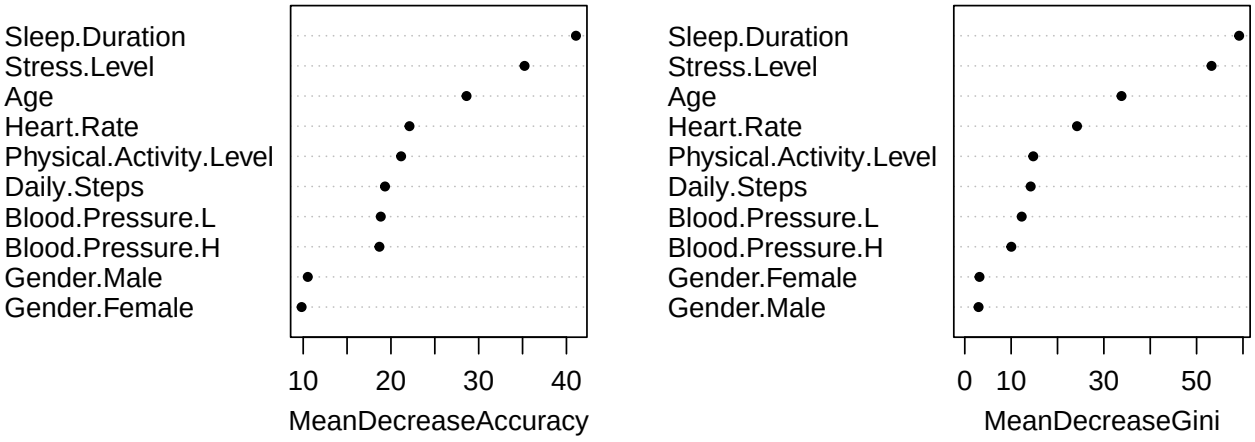
The variable importance plot confirms that, as seen in our single tree, sleep duration, stress level, age and heart rate are the most important variables when estimating the quality of sleep.

We now consider a random forest.

```
##
## yhat.rf  4  5  6  7  8  9
##       4  0  0  0  0  0  0
##       5  0  0  1  0  0  0
##       6  0  0 21  0  0  0
##       7  0  0  0 15  0  0
##       8  0  0  1  0 21  0
##       9  0  0  0  0  0 14
```

```
## [1] 0.04109589
```

r.QS.rf



```
##                           Model        MAE
## 1             Baseline Average 0.98630137
## 2 Baseline Occupation Average 0.56164384
## 3            Linear Regression 0.09589041
## 4                Random Forest 0.04109589
```

Our random forest gives a better MAE than bagging or a single tree. Our graph of variable importance is more balanced as we do not consider all variable at each node but a random subset; that being said, we once again find the same 4 variables at the top with the addition of 'physical activity level'.

Summary

This work was done on a realistic synthetic data set that allows us to put in practice statistical methods to analyse data related to the quality of sleep. We discussed and tackled the ordered classification trough different angles: regression, classification. There exist other ways to handle this problem. We caught relationships between the quality of sleep and the following features: sleep duration, stress level, age, daily steps. Which means that the quality of sleep depends on over variables on which we play a role, trough physical activities. The response was by definition "subjective" but we managed to implement models that can predict this scale of quality.