Sprawozdanie

z projektu w ramach Statistical Learning w praktyce

Agnieszka Tracz
Piotr Lachowicz
Marcin Krzemiński

styczeń 2024

W ramach naszego projektu przygotowaliśmy analizę zbioru danych dotyczących efektywności snu. Dane pochodzą ze strony: https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data i zostały zebrane od 452 uczestników w wieku od 9 do 69 lat; 224 kobiet i 228 mężczyzn.

Opis zbioru danych:

Na nasz zbiór danych składają się 452 obserwacje z danymi w 15 kolumnach:

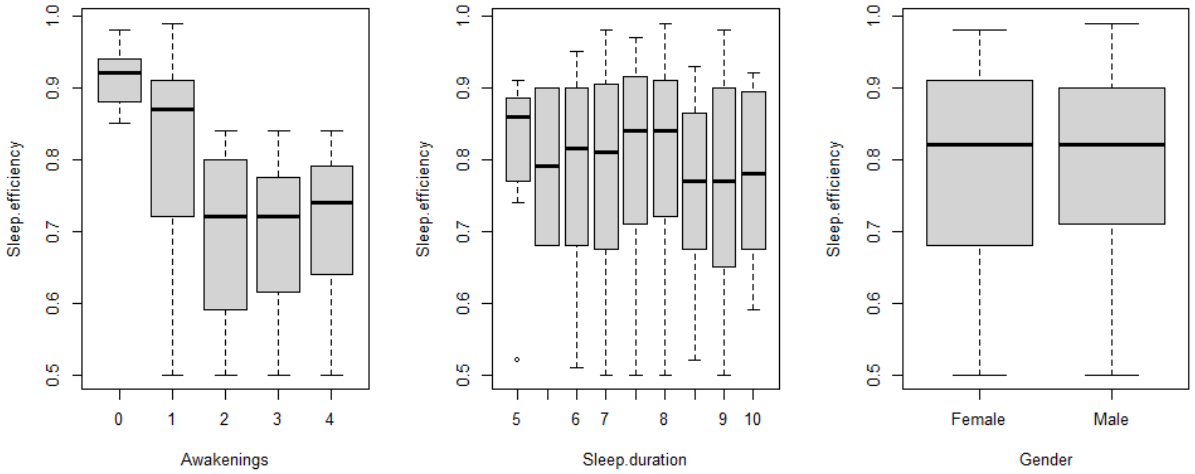
- ID - unikalny identyfikator każdego podmiotu badania
- Age – wiek osoby badanej
- Gender – płeć
- Bedtime – pora o której badany kładzie się spać każdej nocy
- Wakeup time – pora o której osoba badana budzi się każdego ranka
- Sleep duration – całkowity czas, przez który badany spał (w godzinach)
- Sleep efficiency - miara proporcji czasu spędzonego w łóżku na spaniu
- REM sleep percentage - procent całkowitego czasu snu spędzonego w fazie REM
- Deep sleep percentage - procent całkowitego czasu snu spędzonego w stanie głębokiego snu
- Light sleep percentage - procent całkowitego czasu snu spędzonego w lekkim śnie
- Awakenings – liczba ile razy badany obudził się w nocy
- Alcohol consumption - ilość alkoholu spożywanego w ciągu 24 godzin przed snem (w Oz)
- Caffeine consumption- Ilość kofeiny spożywanej w ciągu 24 godzin przed snem (w mg)
- Smoking status – czy badany pali
- Exercise frequency – ile razy badany ćwiczy na tydzień

Cel projektu:

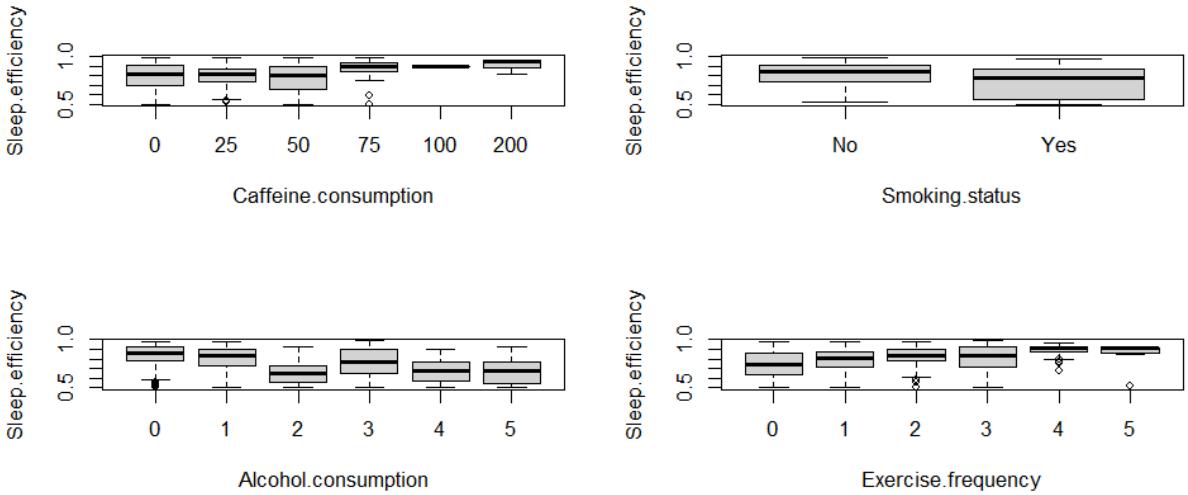
Celem naszej analizy będzie stworzenie kilku modeli predykcyjnych, które na podstawie zebranych danych będą prognozować efektywność snu nowych osobników.

Wstępna analiza danych:

Zastanowimy się jak poszczególne czynniki wpływają na efektywność snu.

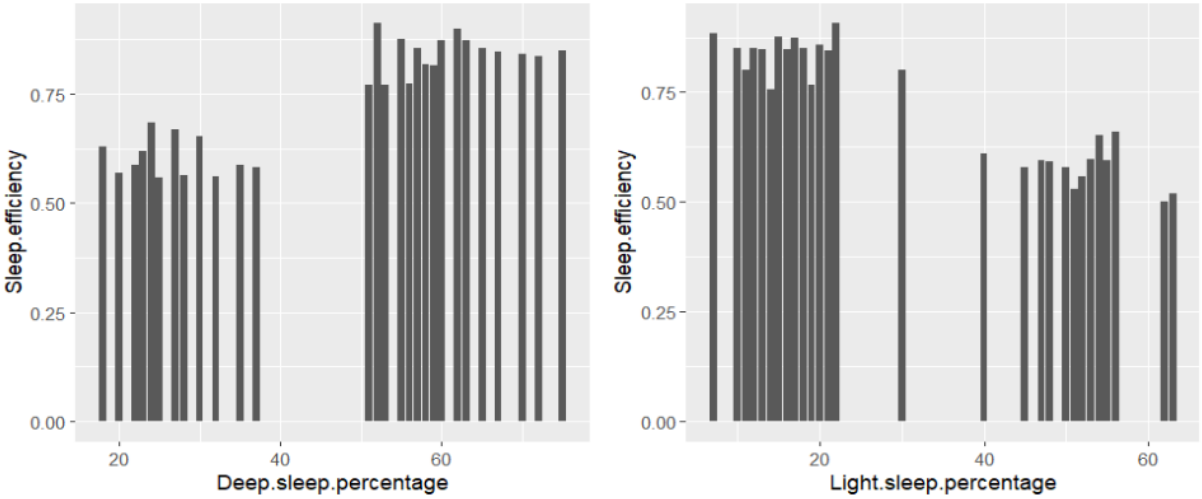


Analizując powyższe wykresy, można zauważyć, że istnieje wyraźna tendencja: im częściej badany budził się w nocy, tym efektywność jego snu była coraz słabsza. To sugeruje, że czynniki związane z nieregularnym snem mogą wpływać negatywnie na jakość snu. Natomiast, brak wyraźnej zależności pomiędzy płcią badanego a efektywnością snu oraz między długością snu a efektywnością snu sugeruje, że te czynniki mogą mieć mniejszy wpływ na ogólną jakość snu w naszym badaniu.



Z wykresów możemy zauważyć, że spożywanie kawy nie powinno negatywnie wpływać na sen. Z drugiej strony, spożywanie alkoholu w większych ilościach wydaje się znacząco obniżać efektywność snu. Dodatkowo, istnieje zauważalna różnica w efektywności snu między palącymi a niepalącymi osobami, gdzie osoby niepalące wykazują nieco wyższą efektywność snu.

Warto również zauważyć, że aktywność fizyczna wydaje się korzystnie wpływać na jakość snu, co potwierdza pozytywny związek między regularną aktywnością fizyczną a efektywnością snu.



Zauważamy, że długość faz snu głębokiego i lekkiego również wpływa na efektywność snu. Pozostałe czynniki nie wydawały się mieć żadnego wpływu na jakość snu.

Przygotowanie danych do analizy:

Sprawdzamy strukturę naszych danych

```
'data.frame':   452 obs. of  15 variables:
 $ ID                    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Age                   : int  65 69 40 40 57 36 27 53 41 11 ...
 $ Gender                : chr  "Female" "Male" "Female" "Female" ...
 $ Bedtime               : chr  "2021-03-06 01:00:00" "2021-12-05 02:00:00" "2021-05-25 21:30:00" "2021-11-03 02:30:00" ...
 $ Wakeup.time           : chr  "2021-03-06 07:00:00" "2021-12-05 09:00:00" "2021-05-25 05:30:00" "2021-11-03 08:30:00" ...
 $ Sleep.duration        : num  6 7 8 6 8 7.5 6 10 6 9 ...
 $ Sleep.efficiency      : num  0.88 0.66 0.89 0.51 0.76 0.9 0.54 0.9 0.79 0.55 ...
 $ REM.sleep.percentage  : int  18 19 20 23 27 23 28 28 28 18 ...
 $ Deep.sleep.percentage : int  70 28 70 25 55 60 25 52 55 37 ...
 $ Light.sleep.percentage: int  12 53 10 52 18 17 47 20 17 45 ...
 $ Awakenings            : num  0 3 1 3 3 0 2 0 3 4 ...
 $ Caffeine.consumption  : num  0 0 0 50 0 NA 50 50 50 0 ...
 $ Alcohol.consumption   : num  0 3 0 5 3 0 0 0 0 0 ...
 $ Smoking.status        : chr  "Yes" "Yes" "No" "Yes" ...
 $ Exercise.frequency    : num  3 3 3 1 3 1 1 3 1 0 ...
```

Zauważamy, że pierwsza kolumna nie jest potrzebna w naszej analizie, ponieważ zawiera jedynie indeksy. W związku z tym usuwamy ją z naszej bazy danych. Podobnie postępujemy z kolumną Wakeup.time, którą także możemy pominąć, ponieważ jest jednoznacznie wyznaczona przez kolumny: Bedtime i Sleep.duration.

Następnie usuwamy wiersze, zawierające niepełne informacje (NA) i zmieniamy typ niektórych zmiennych na „factor” (zmienne jakościowe): Gender, Bedtime, Awakenings, Exercise.frequency oraz Smoking.status.

Ostatecznie otrzymujemy bazę danych złożoną z 388 obserwacji i 13 cech (licząc ze zmienną objaśnianą).

```
      Age           Gender       Bedtime    Sleep.duration   Sleep.efficiency REM.sleep.percentage Deep.sleep.percentage
 Min.   : 9.00   Female:194   00:00  : 64   Min.   : 5.000   Min.   :50.00    Min.   :15.00        Min.   :18.00
 1st Qu.:29.00   Male  :194   22:00  : 51   1st Qu.: 7.000   1st Qu.:70.00    1st Qu.:20.00        1st Qu.:51.00
 Median :41.00                23:00  : 42   Median : 7.500   Median :82.00    Median :22.00        Median :58.00
 Mean   :40.83                01:00  : 35   Mean   : 7.451   Mean   :78.93    Mean   :22.68        Mean   :52.82
 3rd Qu.:52.00                21:00  : 30   3rd Qu.: 8.000   3rd Qu.:90.00    3rd Qu.:25.00        3rd Qu.:63.00
 Max.   :69.00                21:30  : 30   Max.   :10.000   Max.   :99.00    Max.   :30.00        Max.   :75.00
                              (Other):136
 Light.sleep.percentage Awakenings Caffeine.consumption Alcohol.consumption Smoking.status Exercise.frequency
 Min.   : 7.0           0: 87      Min.   :  0.00       Min.   :0.000       No :255        0:110
 1st Qu.:15.0           1:141      1st Qu.:  0.00       1st Qu.:0.000       Yes:133        1: 78
 Median :18.0           2: 49      Median :  0.00       Median :0.000                      2: 45
 Mean   :24.5           3: 55      Mean   : 22.68       Mean   :1.147                      3:113
 3rd Qu.:24.0           4: 56      3rd Qu.: 50.00       3rd Qu.:2.000                      4: 35
 Max.   :63.0                      Max.   :200.00       Max.   :5.000                      5:  7
```

Właściwa analiza danych:

Do predykcji efektywności snu, która jest u nas zmienną ilościową, zbudowaliśmy kilka modeli regresyjnych, zarówno liniowych jak i drzewiastych. Przy modelach liniowych wykorzystywaliśmy metody wyboru podzbioru zmiennych oraz redukcji wymiaru. Jakość modelu ocenialiśmy na podstawie pierwiastka z błędu średniokwadratowego (RMSE) na zbiorze testowym.

Regresje:

- Selekcja krokowa z użyciem metody walidacji krzyżowej

Wykorzystując 10-krotną walidację krzyżową najmniejszy błąd otrzymaliśmy dla modelu liniowego z 16 zmiennymi, wśród których znalazły się te dotyczące palenia, ćwiczeń fizycznych, liczby przebudzeń, godziny pójścia spać, snu lekkiego, wieku oraz spożycia alkoholu i kofeiny. Błąd RMSE na zbiorze testowym wyniósł **6,72**.

- Regresja liniowa (MNK)

Dla modelu liniowego współczynniki przy cechach takich jak: płeć, długość snu czy spożycie kofeiny okazywały nieistotnie (p-wartość>5%) różne od zera. Błąd RMSE na zbiorze testowym wyniósł **6,13**.

- Regresja grzbietowa

Parametr kary *lambda* dobrany metodą walidacji krzyżowej wyszedł równy 1,20. Dla tego parametru błąd RMSE na zbiorze testowym wyniósł **6,11**.

- Regresja lasso

Podobnie jak w przypadku regresji grzbietowej parametr kary *lambda* został dobrany metodą walidacji krzyżowej. Optymalna jego wartość wyniosła 0,20, a model z tym parametrem wykazywał na teście błąd RMSE rzędu **6,12**. Wyzerowaniu uległy nieliczne współczynniki (głównie te odpowiedzialne za płeć i godzinę pójścia spać).

- PCR

Wybierając redukcję z najmniejszym błędem, metoda ta pozwoliła nam zmiejszyć wymiar z 28 jedynie o 3 „poziomy”, do wymiaru 25, dla którego błąd RMSE wyniósł **6,08**. Model z 25 zmiennymi (zbudowany na pełny zbiorze danych) wyjaśnia 84% wariancji efektywności snu.

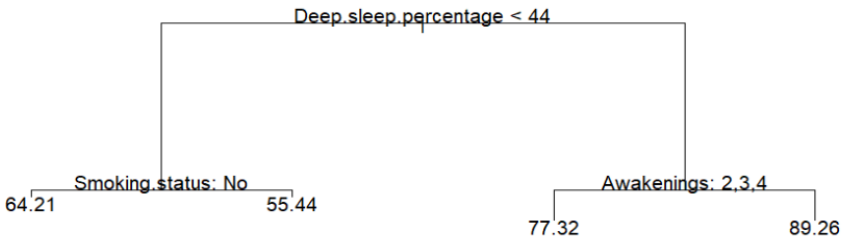
- PLS

Metoda ta pozwala nam zredukować wymiar z 28 do 3 (sic!) przy zachowaniu tego samego poziomu wyjaśnialności wariancji co w PCR (84%) oraz obniżeniu błędu testowego RMSE do wartości **6,01**.

Modele drzewiaste:

- Drzewo decyzyjne

Ze względu na dużą interpretowalność, pierwszym zastosowanym przez nas modelem z metod drzewiastych jest drzewo decyzyjne.



Zauważamy, że w modelu zostały uwzględnione tylko trzy czynniki, te same, które we wstępnej analizie uznaliśmy za wpływowe na jakość snu. Mimo prostoty modelu wykazuje on większą wydajność niż metody regresyjny. Błąd RMSE na zbiorze testowym wyniósł w tym przypadku **5,26**. Drzewa przycięte do mniejszej liczby liści wykazują istotnie większe błędy predykcji.

- Bagging

Modele typu bagging rozważaliśmy z różną liczbą drzew. Najlepsze wyniki na poziomie błędu RMSE równym **4,80** osiągnęliśmy dla parametru *ntree*=225 (liczba drzew).

- Random Forest

Również w przypadku lasu losowego rozważaliśmy różne wartości parametrów *mtry* oraz *ntree*. Budując modele z wartościami tych parametrów wziętymi z ustalonej przez nas siatki i porównując ich RMSE, jako najlepszy model wybraliśmy ten z parametrami *mtry*=6 oraz *ntree*=50, dla którego błąd wyniósł **4,59**.

- Boosting

Dla modelu boostingowego rozważaliśmy modele z różnymi wartościami parametrów *shrink* oraz *interaction.depth*. Dla każdego z nich liczba drzew użytych do predykcji była dobierana metodą walidacji krzyżowej. Najmniejszy błąd RMSE równy **4,94** osiągnął model z parametrami *shrink*=0.01 oraz *interaction.depth*=5.

- BART

Budując domyślny model BART otrzymaliśmy dla zbioru testowego błąd RMSE na poziomie **5,22**.

- XGBoost

Rozważając model xgb z parametrem *max.depth*=3, najmniejszy błąd RMSE na poziomie **4,66** otrzymaliśmy przy 29 iteracjach algorytmu.

Podsumowanie

Zbudowaliśmy dwanaście modeli mających za zadanie przewidywać efektywność snu na podstawie 12 wyżej wspomnianych cech. Potwierdziły one nasze przypuszczenia ze wstępnej analizy, pokazując, że na efektywność snu najbardziej wpływają takie czynniki, jak: liczba przebudzeń, długość poszczególnych faz snu, czy bycie palaczem. Z kolei mniej istotna jest płeć osoby badanej.

Wedle przyjętego przez nas kryterium oceny modelu metody drzewiaste poradziły sobie lepiej (mniejszy błąd) z postawionym zadaniem. Chcąc dokonywać predykcji dla nowych obserwacji, najlepiej skorzystać z modelu opartego na Random Forest lub XGBoost.