Final project of the course

Mathematics for Computer Science

Dmytro Gavrylchenko

- 38 years
- Senior PM
- Can't live without Python

Data analysis and preparation

Sleep Efficiency Dataset from Kaggle

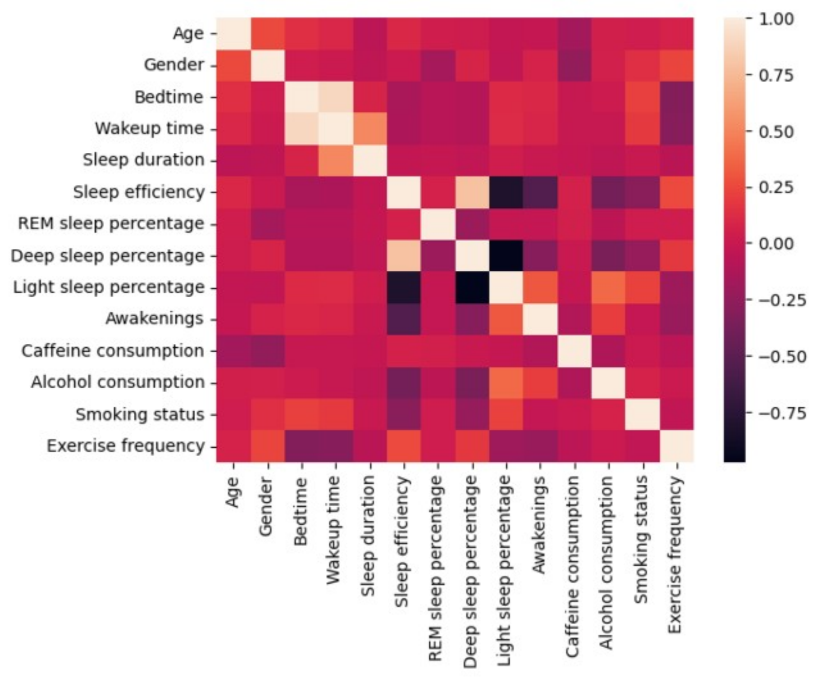
The dataset contains information on a group of 452 subjects and their sleep patterns. Each test subject is identified by a unique "subject identifier" and their age and gender are recorded. The Sleep Time and Wake Time fields indicate when each subject slept and woke up each day, and the Sleep Duration field records the total amount of time each subject slept in hours. The "Sleep efficiency" field is an indicator of the proportion of time spent in bed from actual sleep.

The REM Sleep Percentage, Deep Sleep Percentage, and Light Sleep Percentage fields show the time each subject spent in each stage of sleep. The Awakening field records the number of times each subject woke up during the night. In addition, the dataset contains information on each subject's caffeine and alcohol consumption in the 24 hours before bedtime, their smoking status, and their exercise frequency.

Data analysis

A surface analysis of the data made it possible to establish collinearity between the features "Bedtime", "Wakeup time", "Sleep duration", as well as between "REM sleep percentage", "Deep sleep percentage", "Light sleep percentage". Redundant columns were removed to avoid overloading the model with homogeneous information. It can also be concluded from the previous analysis that the quality of sleep depends most on the duration of deep sleep and to a lesser extent on the duration of awakening.

	Feature	VIF
0	Age	1.152335
1	Gender	1.267567
2	Bedtime	inf
3	Wakeup time	inf
4	Sleep duration	inf
5	Sleep efficiency	5.101836
6	REM sleep percentage	48.718475
7	Deep sleep percentage	113.932769
8	Light sleep percentage	18.323645
9	Awakenings	1.751138
10	Caffeine consumption	1.105847
11	Alcohol consumption	1.229921
12	Smoking status	1.271401
13	Exercise frequency	1.286044



Data preparation

- ❑ Among the data, 65 NaNs (1% of all features) were observed, which were replaced by column averages;
- ❑ Male/Female and Yes/No binaries are replaced by 1/0;
- ❑ The time from DateTime format is replaced by float and the start of the day is shifted to 21:00.

ID	Age	Gender	Bedtime	Wakeup time	Sleep duration	Sleep efficiency	REM sleep percentage	Deep sleep percentage	Light sleep percentage	Awakenings	Caffeine consumption	Alcohol consumption	Smoking status	Exercise frequency
449	40	Female	2021-09-07 23:00:00	2021-09-07 07:30:00	8.5	0.55	20	32	45	1.0	NaN	3.0	Yes	0.0
450	45	Male	2021-07-29 21:00:00	2021-07-29 04:00:00	7.0	0.76	18	72	10	3.0	0.0	0.0	No	3.0
449	**40**	**0**	**2.0**	**10.5**	**8.5**	**0.55**	**20**	**32**	**45**	**1.0**	**23.65**	**3.0**	**1**	**0.0**
450	**45**	**1**	**0.0**	**7.0**	**7.0**	**0.76**	**18**	**72**	**10**	**3.0**	**0.0**	**0.0**	**0**	**3.0**

GO IT Neoversity

Modeling

The purpose of this study is to build a model for predicting sleep quality based on a set of sleep efficiency data from Kaggle. To solve the problem, three models were created using the open Python library - Scikit-Learn:

- Linear Regressor;
- K-Nearest Neighbors Regressor;
- Random Forest Regressor.

Description of models

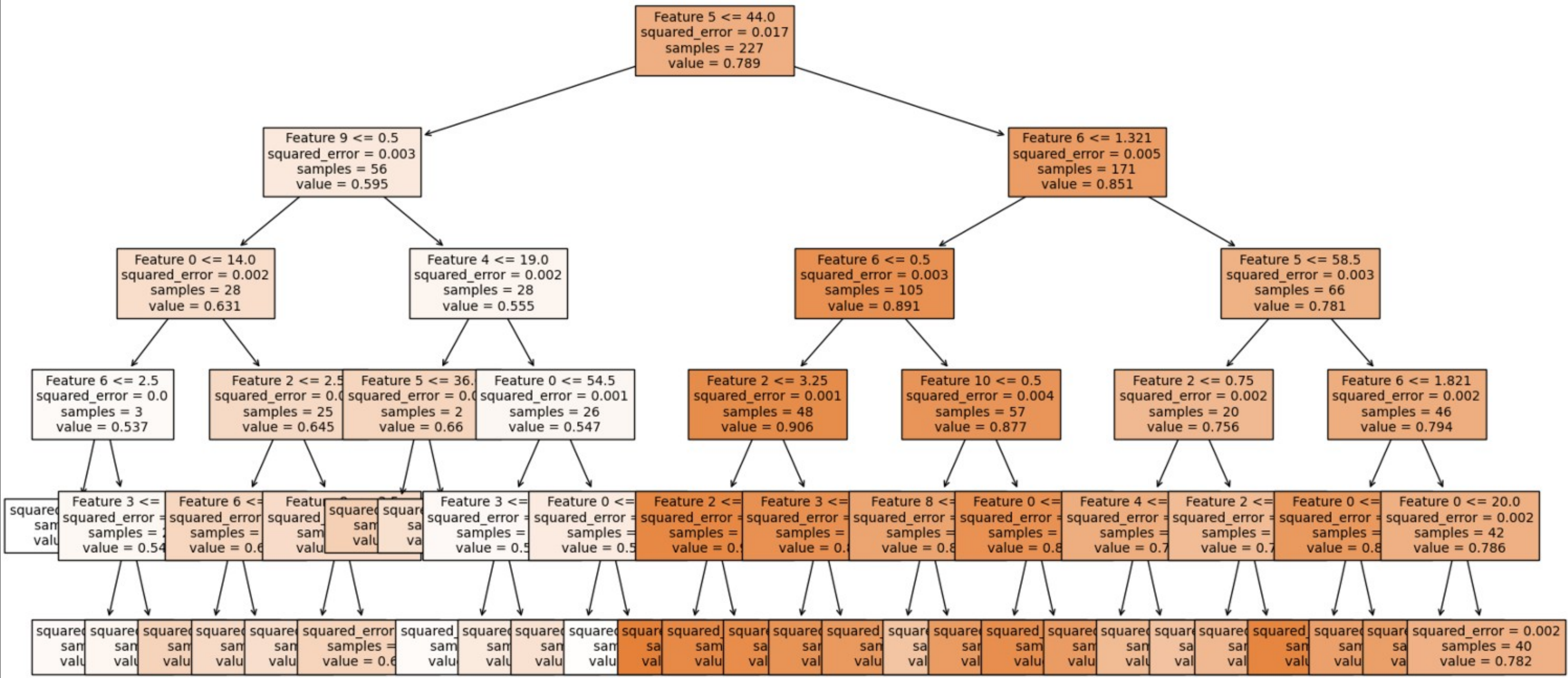
In the process of learning and selecting hyperparameters, the model acquired the following parameters:

- <u>Linear Regressor</u>:
- Coefficients: 0.0005457   0.00890756 0.00090071 0.00155916 0.00704486 0.00554475 -0.03501723 0.0001692 -0.00603226 -0.04709852 0.00362632
- Interception: 0.3667832

- <u>K-Nearest Neighbors Regressor</u>:
- Number of neighbors (k): 5
- Weight function: uniform
- Distance metric: minkowski

- <u>Random Forest Regressor</u>:
- see next slide.

Random Forest Regression - Decision Tree Visualization

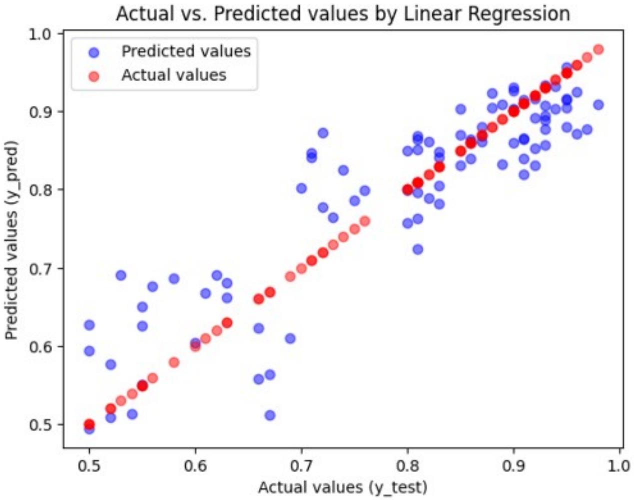


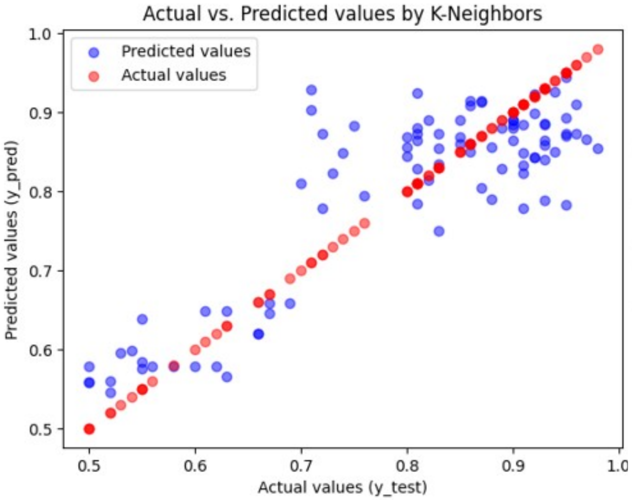
Metrics

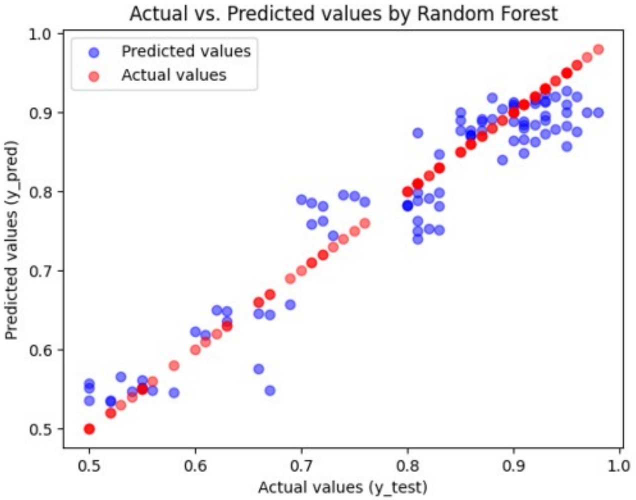
Metrics	Linear Regression	K(5)-Nearest Neighbors	Random Forest
Mean Absolute Error (MAE)	0.04949	0.05732	0.03517
Mean Squared Error (MSE)	0.00400	0.00512	0.00187
Root Mean Square Error (RMSE)	0.06322	0.07158	0.04320
R-squared	0.79425	0.73618	0.90390

Analysis of results

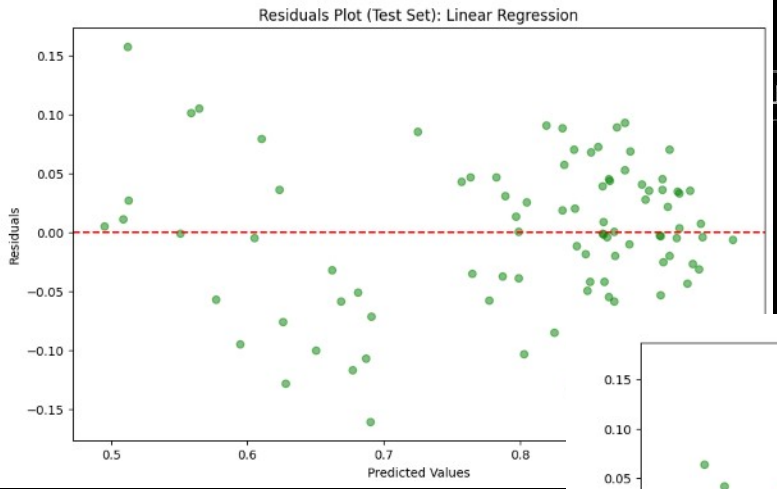
Comparison of actual and predicted data

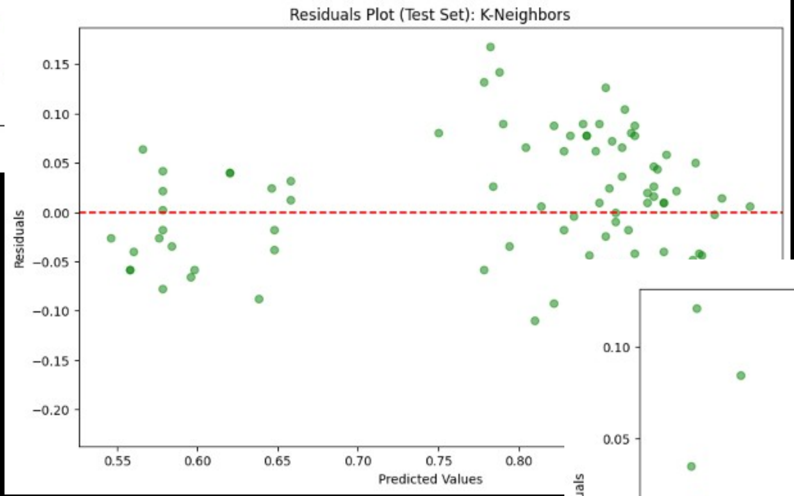


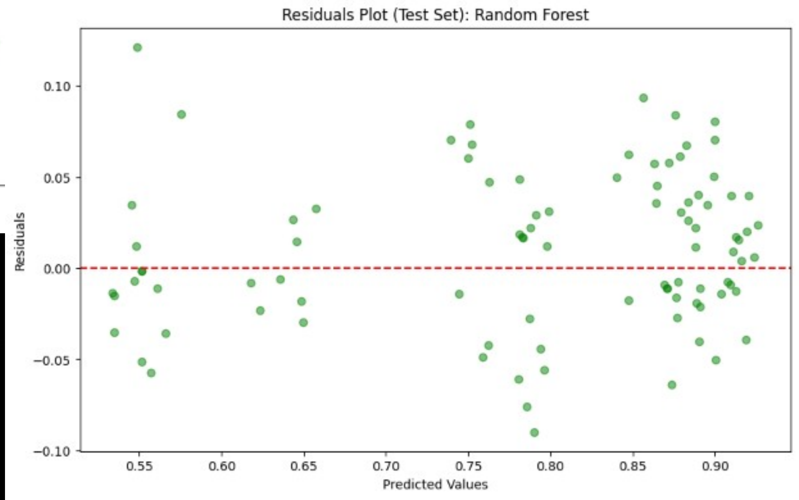




Error analysis







It is easy to see that the random forest based model produces less variation in the residuals due to better prediction. It is interesting that if the residuals of the linear regression are scattered fairly evenly, then the residuals of the random forest are grouped into clouds. This is because this algorithm sorts the values into separate bins and tries to assign predictions to one of the bins it knows, whereas the linear model is a regular polynomial for which the algorithm chooses coefficients.

GO IT Neoversity

Comparison of the quality of models

This table compares the relative deviations in % between the mean of the 10 best and worst predictions. Features marked in orange are those that differ greatly between good and bad predictions and are therefore poorly captured by the model.

Feature	Linear Regression	K(5)-Nearest Neighbors	Random Forest
Age	3.7	27.9	30.0
Gender	25.0	20.0	100.0
Bedtime	30.8	20.6	30.6
Sleep duration	3.3	-2.1	-2.0
REM sleep percentage	-4.0	-0.9	2.2
Deep sleep percentage	-37.2	5.8	17.4
Awakening	75.0	148.7	72.7
Caffeine consumptions	-40.0	-36.9	-20.1
Alcohol consumption	200	-41.7	-5.9
Smoking status	0.0	-100.0	-20.0
Exercise frequency	-23.1	-22.7	38.5
Residuals	6454	2373	1380

Conclusions

Based on the work carried out, it can be established that the model based on the random forest regressor coped best with the task of predicting the quality of sleep. It has both the lowest error (-40% MAE and RMSE compared to the linear regressor and -60% compared to K nearest neighbors) and the smallest variation from best to worst prediction.

Insight

In the process of performing the task, an attempt was made to increase the accuracy of the random forest regressor using the GridSearch algorithm. The model proposed by the algorithm was more complex than the model with standard hyperparameters. However, a more complex model performs worse, which may mean that increasing the complexity does not always lead to better predictions.

Metrics	Random Forest GridSearch	Random Forest Default
Mean Absolute Error (MAE)	0.03892	0.03517
Mean Squared Error (MSE)	0.00231	0.00187
Root Mean Square Error (RMSE)	0.04803	0.04320
R-squared	0.88124	0.90390

Thank you for attention