Stock Price Prediction Methods

Márton Espán

2024. 05. 09.

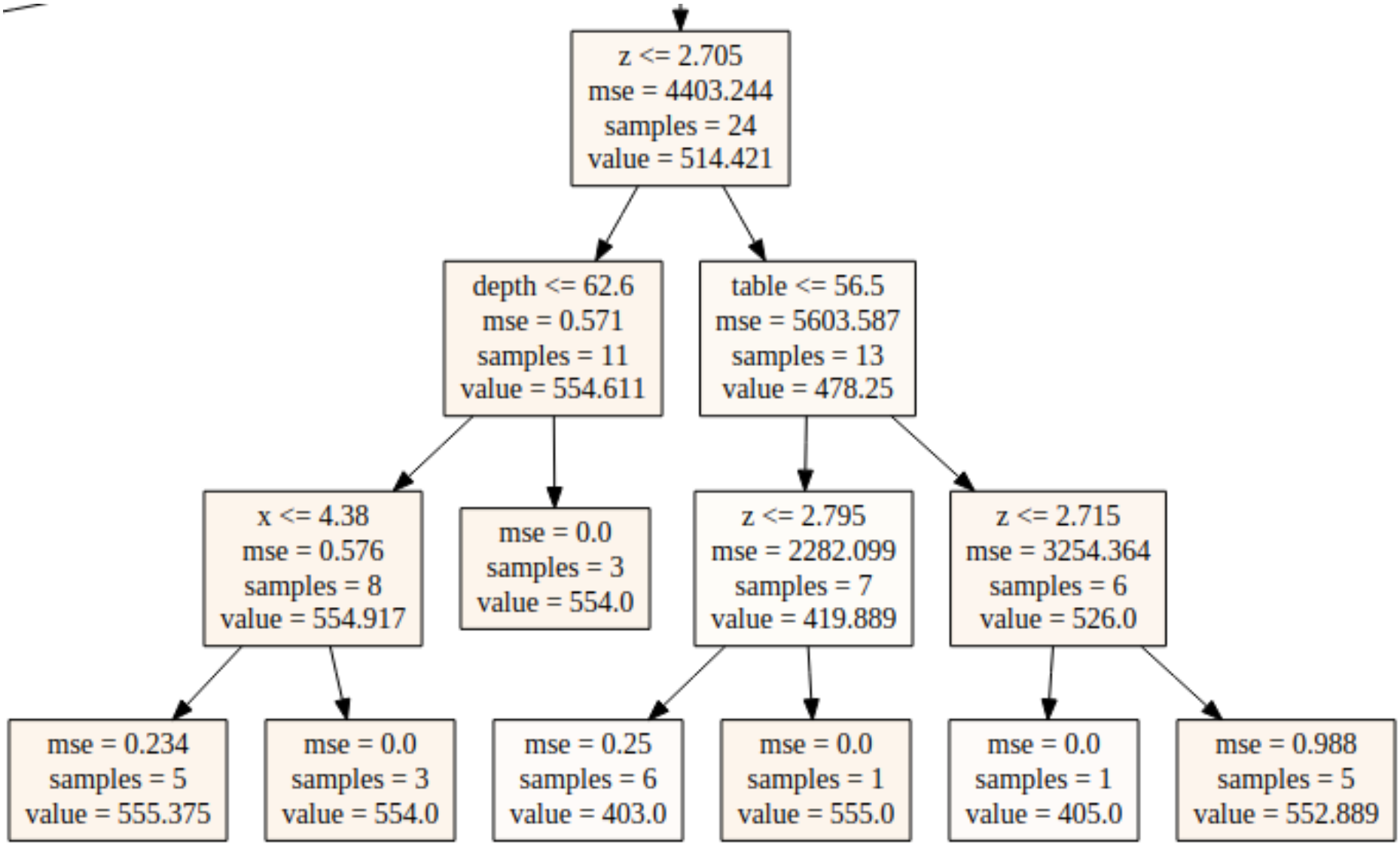
Data Used and Forecasting Procedure

- Yahoo Finance
- S&P 500
- Data from 2013–2023
- Daily closing prices and volume
- If data was missing, the stock was excluded
- 448 remained
- 10 years training (2013–2022), 1 year testing (2023)
- Benchmark: market return and today's stock price

Methods Examined

- Random forest regression
- KNN regression
- AR(1) model
- ARIMA models
- LSTM

Random Forest Regression



RF1 I.

Indicators for RF1:

- Ret-1
- EMA 14, 30, 50, 200
- RSI 14, 30, 50, 200
- PVT
- Volume-1

$$RSI = 100 - \frac{100}{1 + \frac{\text{average gain}}{\text{average loss}}}$$

RF1 II.

Optimized parameters:

- n_estimators= 20
- max_depth= 4
- min_samples_split= 20
- min_samples_leaf= 4

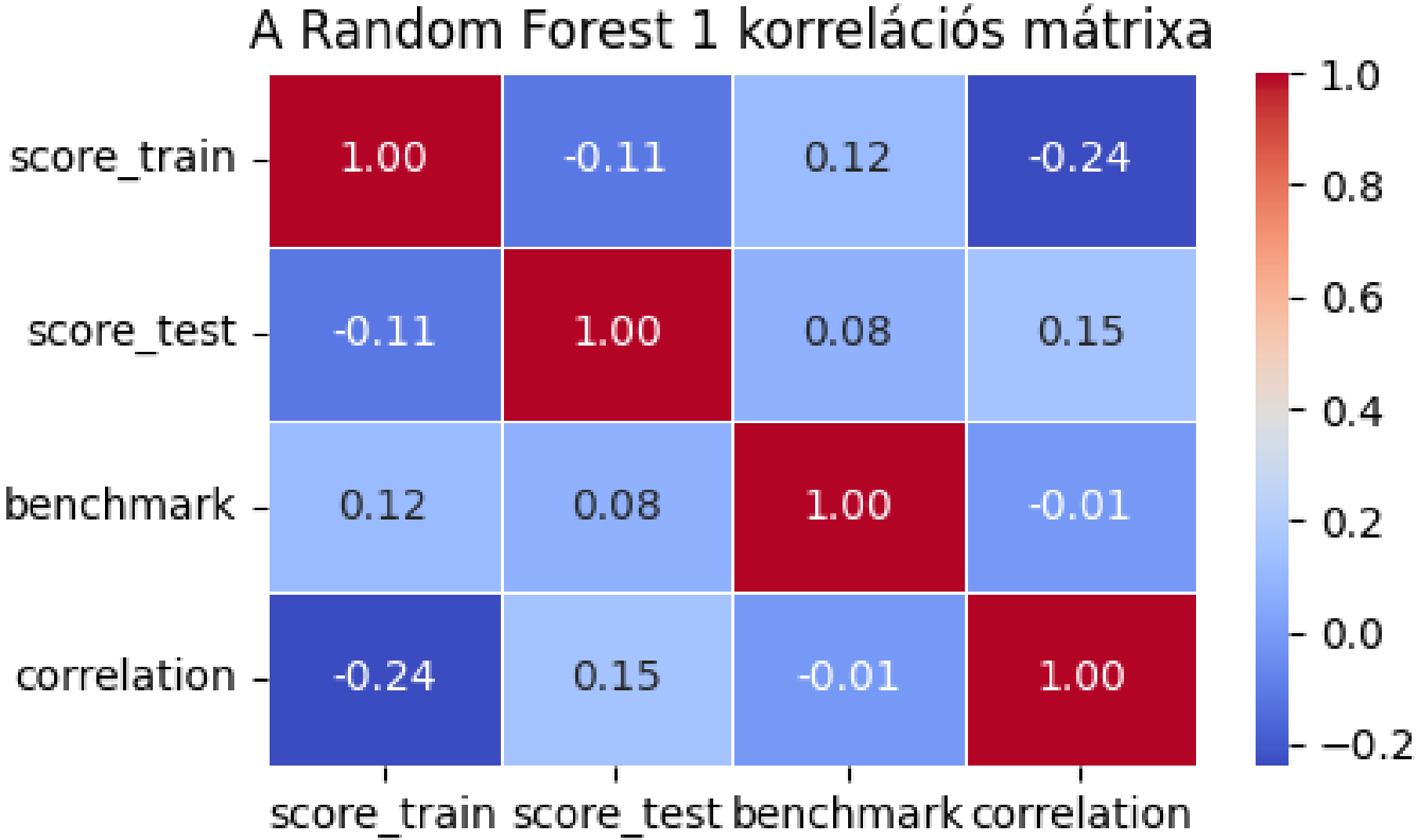
RF1 III.

Results:

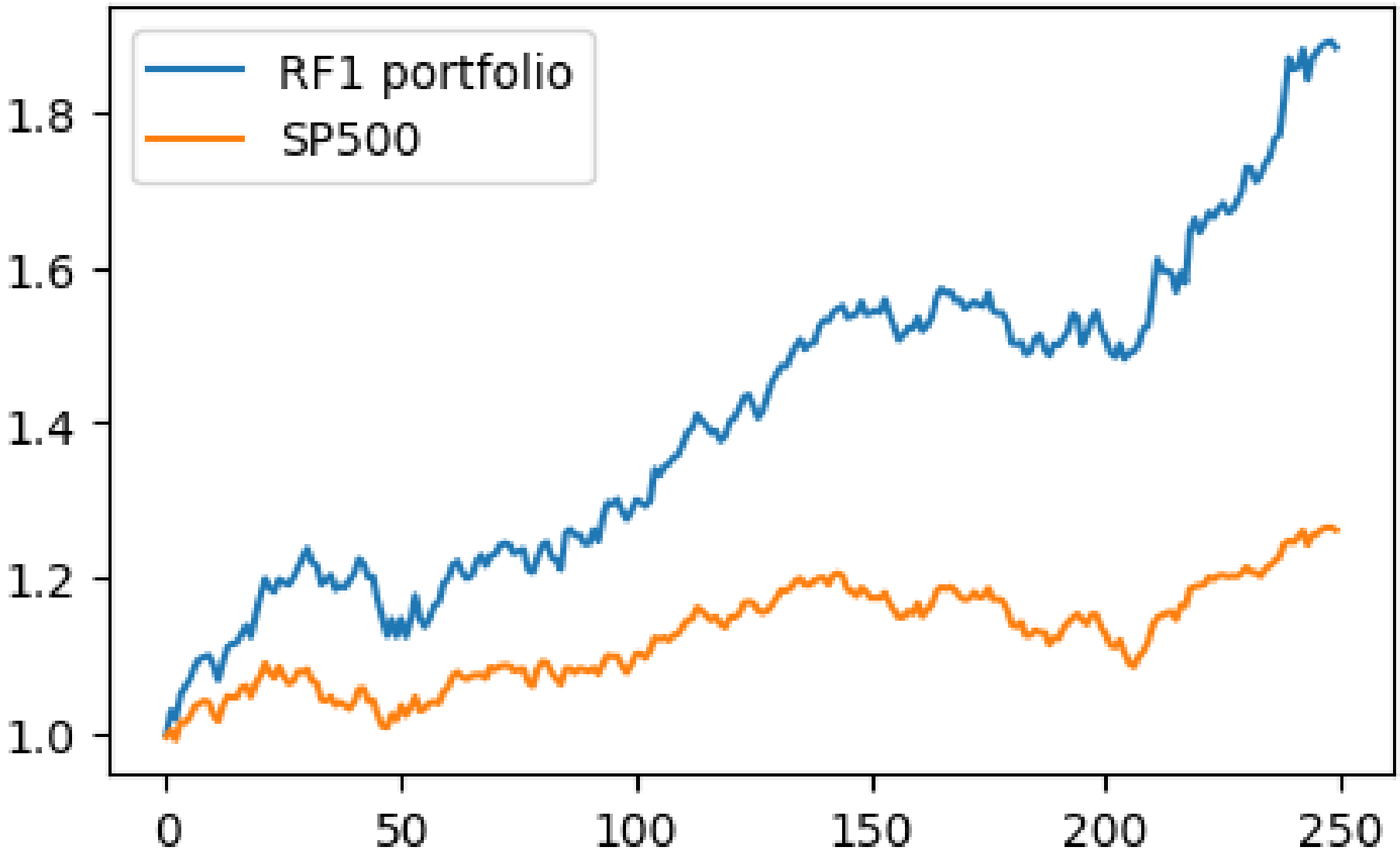
- score_train 0.085606
- score_test -0.057209
- benchmark -0.004352
- correlation 0.040853

$$R^2 = 1 - \frac{\sum(y_i - y_{pred_i})^2}{\sum(y_i - y_{mean})^2}$$

RF1 IV.


A Random Forest 1 korrelációs mátrixa
score_train
score_test
benchmark
correlation
1.00
-0.11
0.12
-0.24
-0.11
1.00
0.08
0.15
0.12
0.08
1.00
-0.01
-0.24
0.15
-0.01
1.00
score_train
score_test
benchmark
correlation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

RF1 V.



RF2 I.

Indicators for RF2:

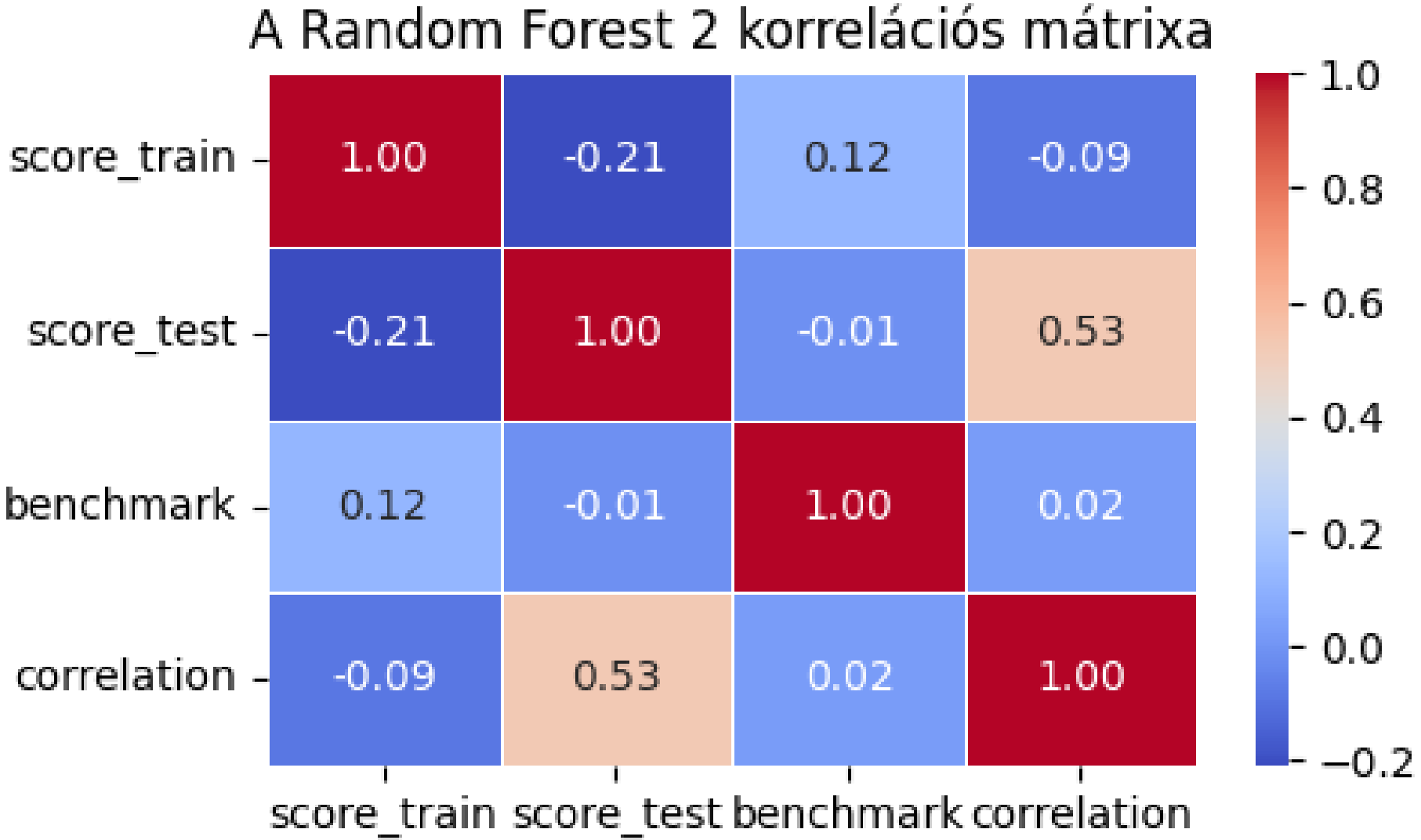
- Ret-1, -2, -3, -4, -5
- Volume-1, -2, -3, -4, -5

RF2 II.

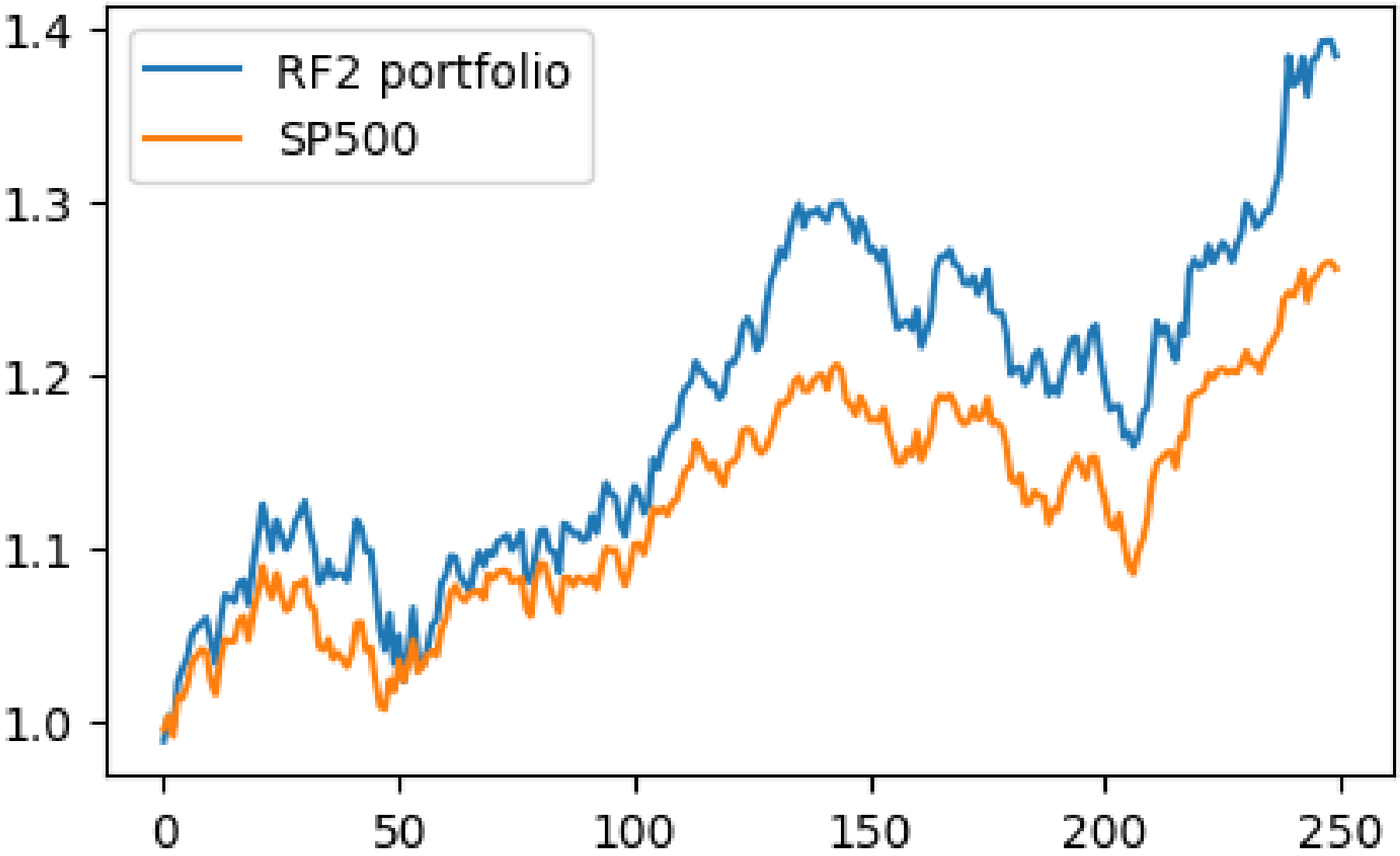
Results:

- score_train 0.078952
- score_test -0.027274
- benchmark -0.004352
- correlation 0.005215

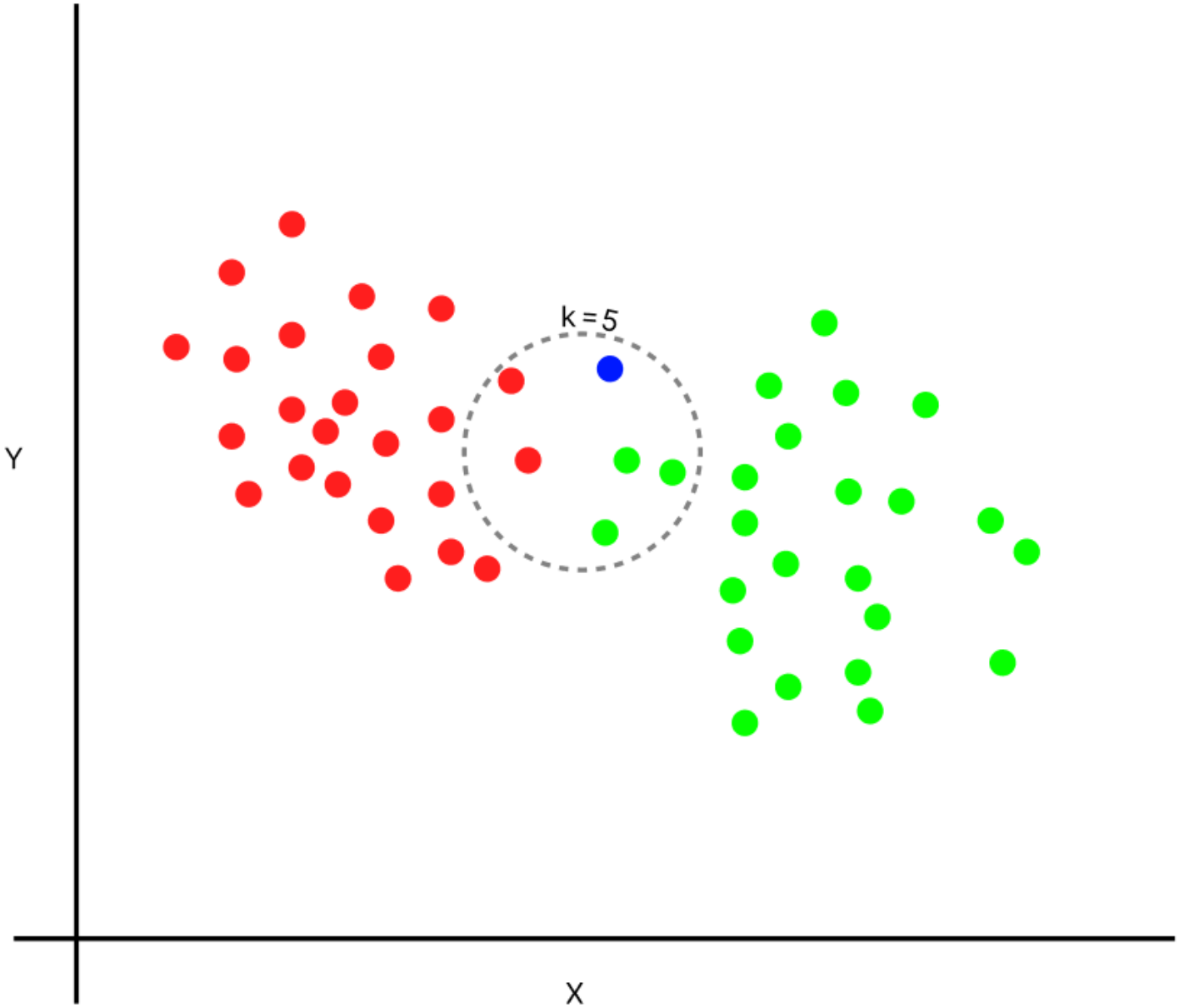
RF2 III.


A Random Forest 2 korrelációs mátrixa
score_train
score_test
benchmark
correlation
1.00
-0.21
0.12
-0.09
-0.21
1.00
-0.01
0.53
0.12
-0.01
1.00
0.02
-0.09
0.53
0.02
1.00
score_train
score_test
benchmark
correlation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

RF2 IV.



KNN Regression

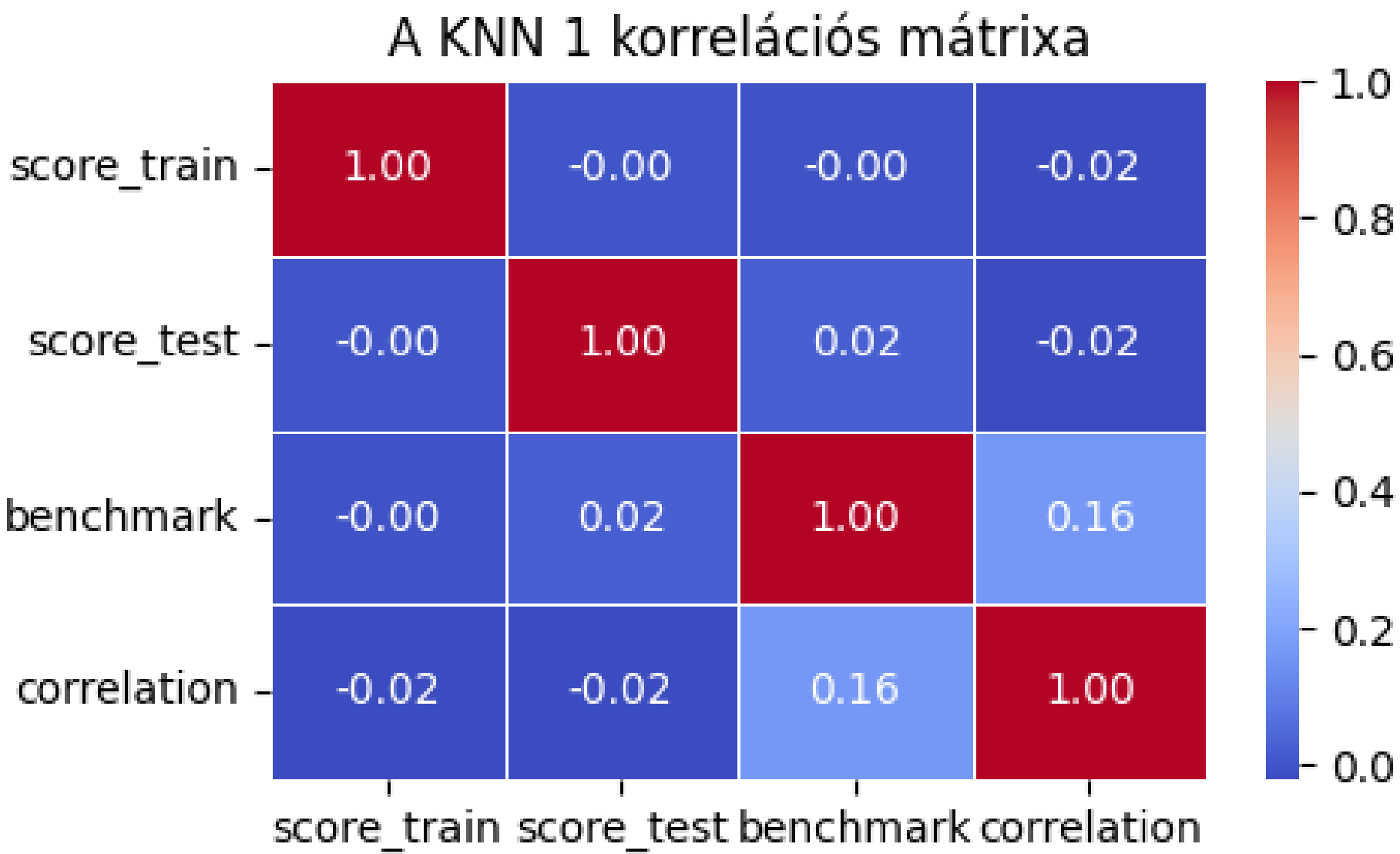


KNN1 I.

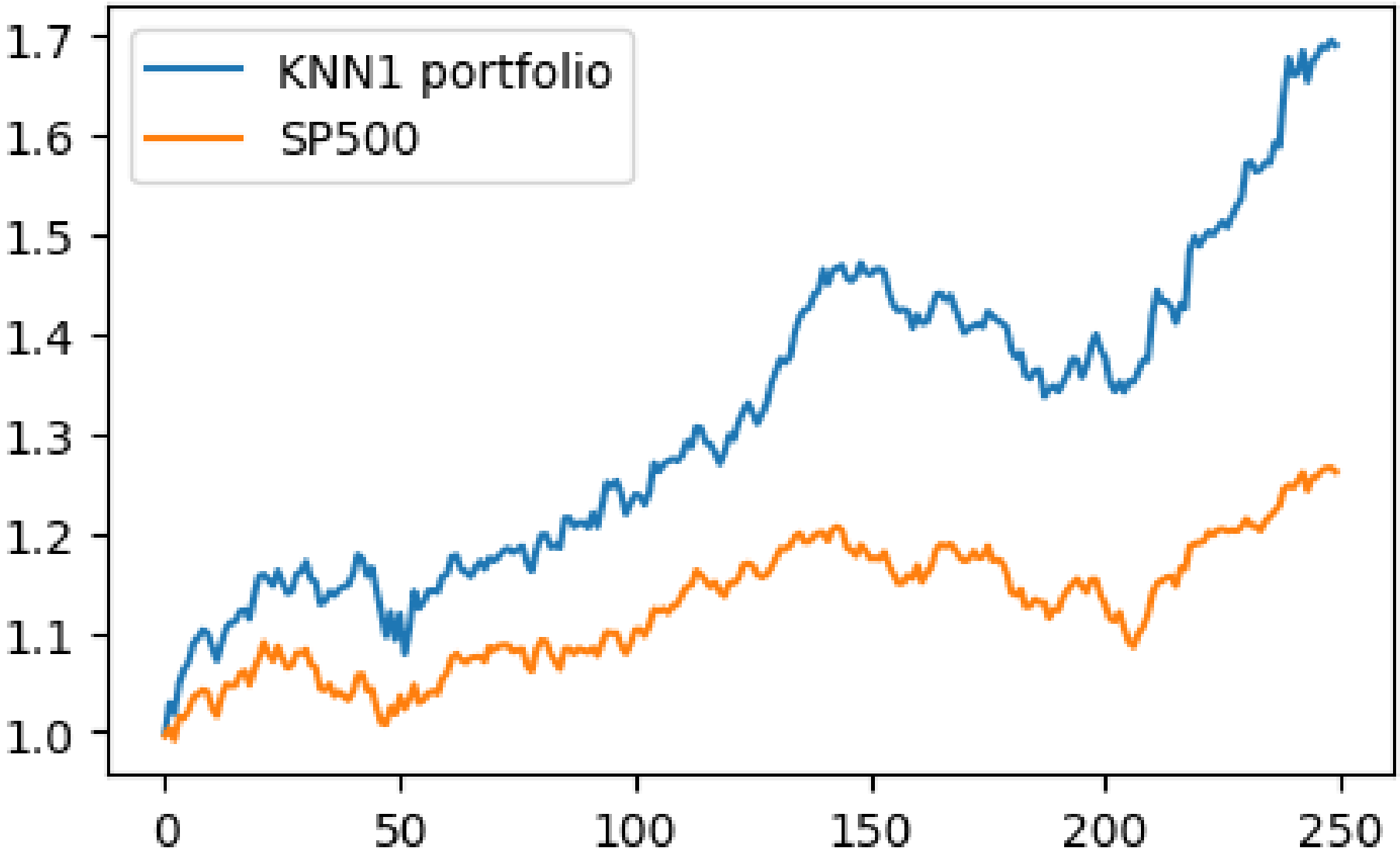
Results:

- score_train 0.113815
- score_test -0.109042
- benchmark -0.004352
- correlation 0.047081

KNN1 II.


A KNN 1 korrelációs mátrixa
score_train
score_test
benchmark
correlation
1.00
-0.00
-0.00
-0.02
-0.00
1.00
0.02
-0.02
-0.00
0.02
1.00
0.16
-0.02
-0.02
0.16
1.00
score_train
score_test
benchmark
correlation
1.0
0.8
0.6
0.4
0.2
0.0

KNN1 III.

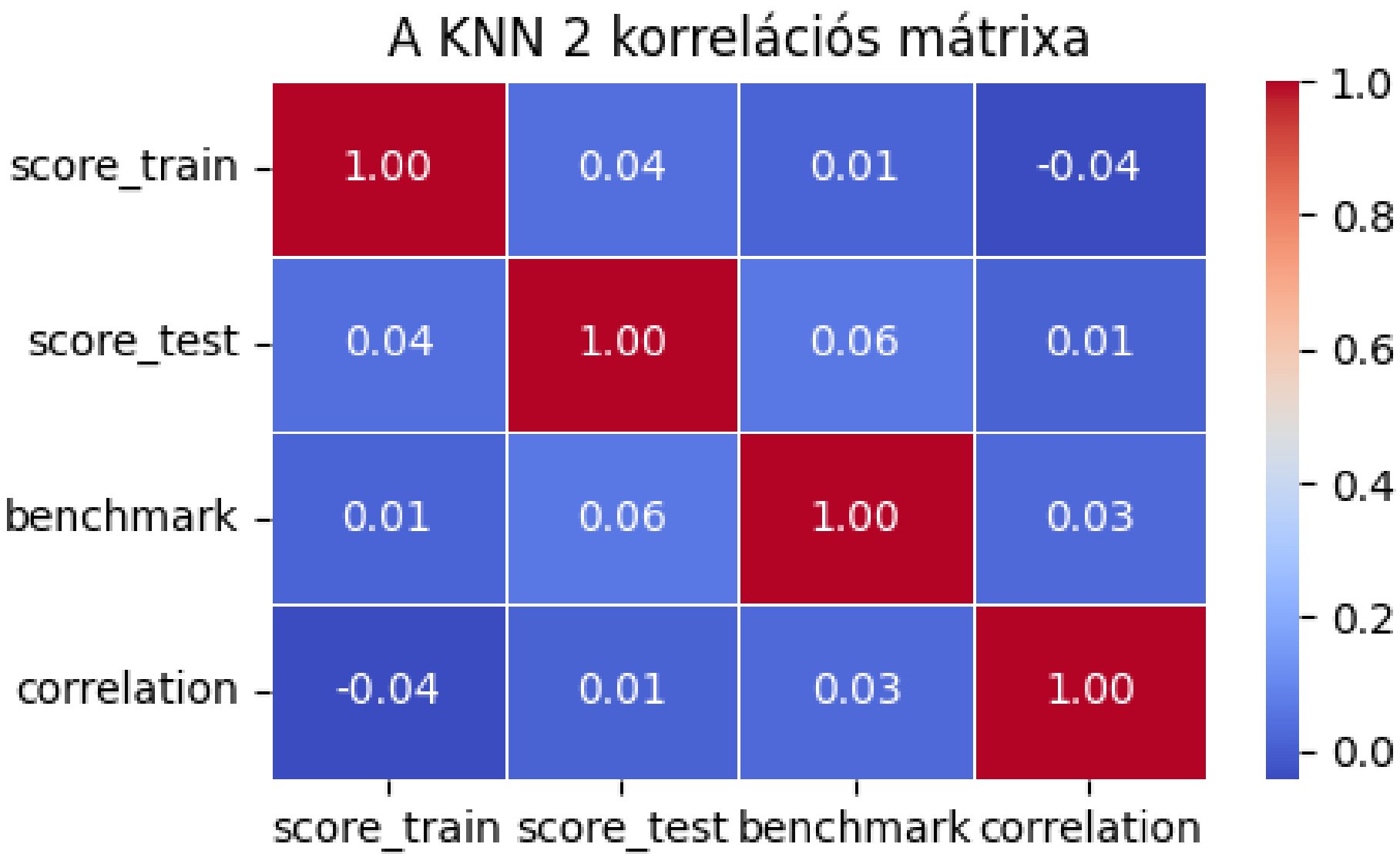


KNN2 I.

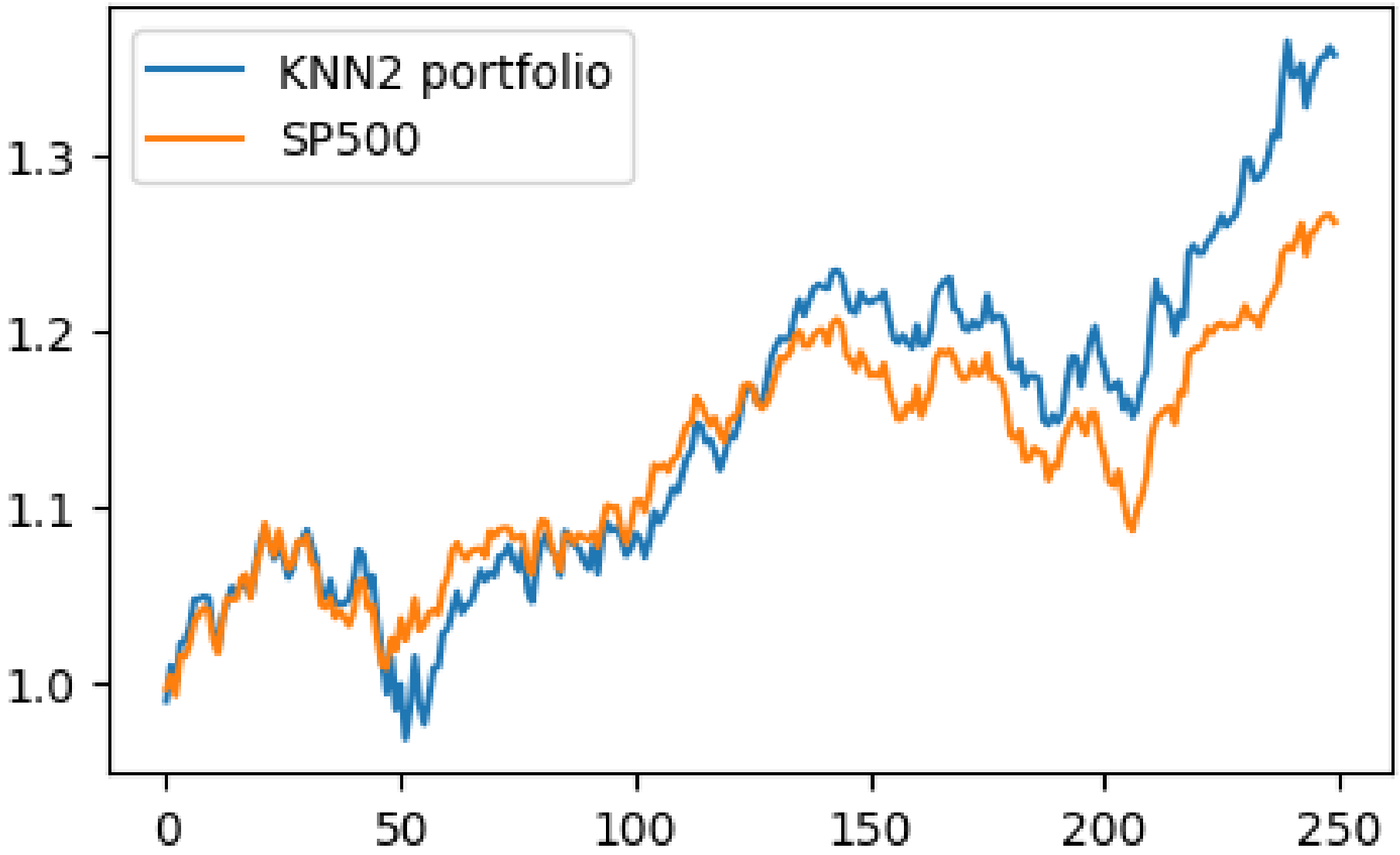
Results:

- score_train 0.061398
- score_test -0.052859
- benchmark -0.004352
- correlation 0.000174

KNN2 II.


A KNN 2 korrelációs mátrixa
score_train
score_test
benchmark
correlation
1.00
0.04
0.01
-0.04
0.04
1.00
0.06
0.01
0.01
0.06
1.00
0.03
-0.04
0.01
0.03
1.00
score_train
score_test
benchmark
correlation
1.0
0.8
0.6
0.4
0.2
0.0

KNN2 III.



ARIMA Models

We say that the time series X_t is an ARMA(p, q) process if

$$X_t = \mu + \sum_{i=1}^{p} \alpha_i \cdot X_{t-i} + \sum_{j=1}^{q} \beta_j \cdot \epsilon_{t-j} + \epsilon_t.$$

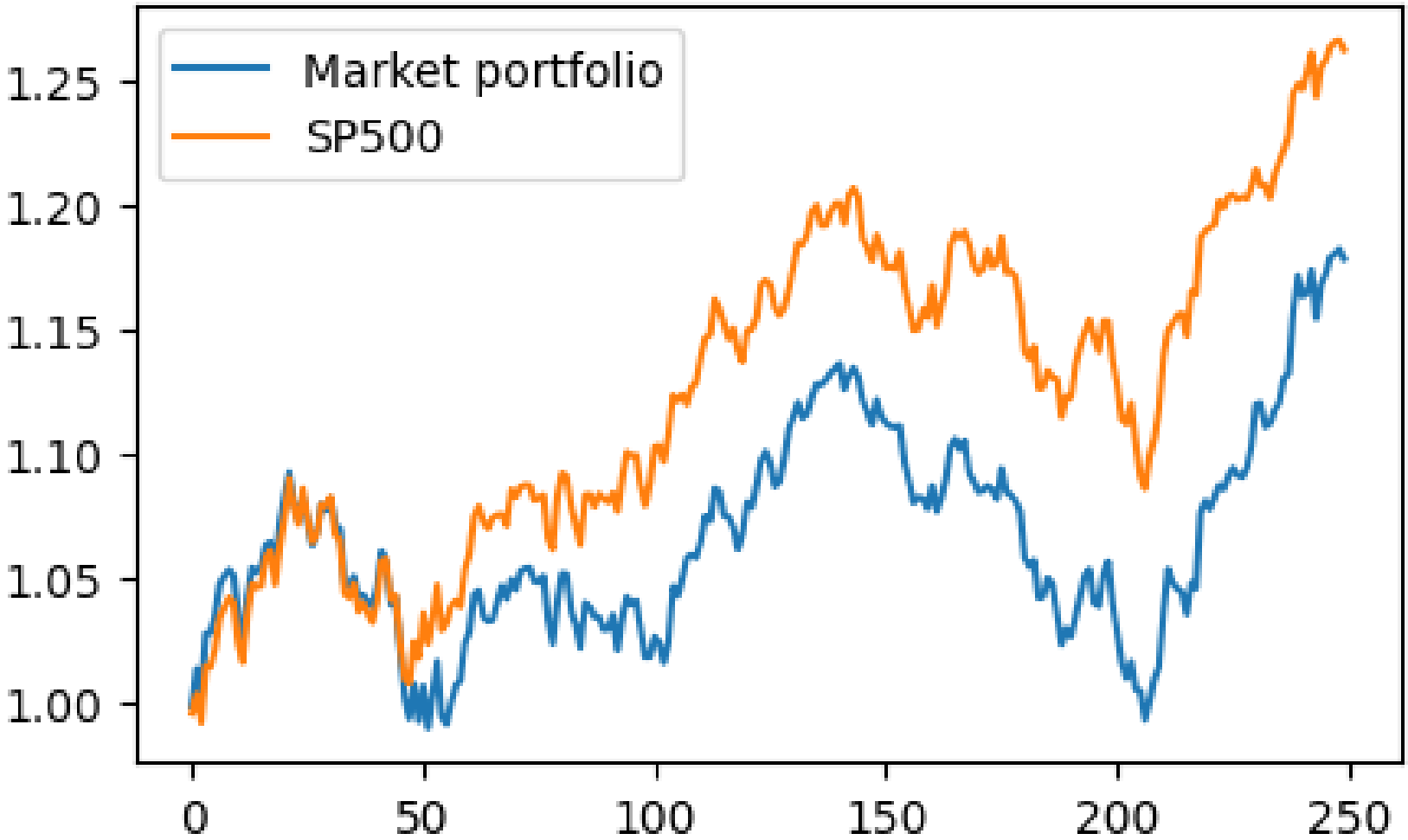
X_t is an ARIMA(p, d, q) process if $X_{t+d} - X_t$ is an ARMA(p, q) process. Thus, if X_t is an AR(1) process, then $X_t = \mu + \alpha_1 \cdot X_{t-1} + \epsilon_t$.

LIN1 I.

Results:

- score_train 0.003966
- score_test -0.005593
- benchmark -0.004352
- p_value 0.186823
- correlation 0.007811

True Benchmark



Thank you for your attention!

Image sources:
[1] https://pub.towardsai.net/...
[2] https://neptune.ai/...
[3] https://medium.com/...