Multivariate Regression on Garmin running data

Michel Mariën (852674168) [1]

1. Introduction

I have been a recreational runner for the past 18 years. I first started running because of a Flemish scientific television called 'Marathon' in which 6 average non-sportive Belgians are coached and supported to run a half marathon in 6 months and a full marathon in 12 months ((Leuven, 10-10-2006)). Due to that program I have been running ever since. I don't have the ambition to run the full marathon in the short term or trying to improve my personal records but I have always been interested in which factors influence my average speed.

For the last four years, I have used a Garmin smartwatch that recorded not only my track times but also the routes that I have run. Since I mostly run to stay in shape, there is no additional for these activities.

There is an article available online that tries to use Garmin Running data to predict finish times for new races (Acosta, 4-4-2021). Unfortunately, the writer never published the second article (Cam, 20-3-2022), so the author's method for further analysis of the data remains unclear. This report builds on top of that first article and seeks to further analyse Garmin running data using python.

2. Goal

For this report, I want to test if the data saved in my Garmin-profile alone is enough to predict average track speeds using a regression model. The use of a regression model as the basis for a prediction model is chosen because of its simplicity and explainability. Next to the main question whether a regression model can be used to predict the average speed on a track, several additional questions are:

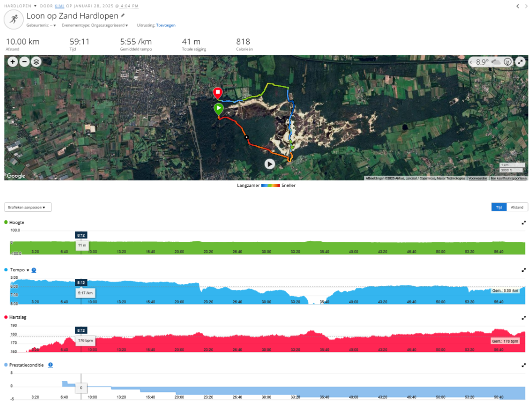
1. Which factors have the biggest influence on the recorded average speed within the dataset
2. Is there a correlation between distance and average speed
3. Is there a correlation between dunes and average speed

This report will consists of 2 data analyses, one analysis based on a direct export of recorded track data and a second analysis based on the direct export and additional features that are also stored in a Garmin account but not exported. These data analyses will be compared to see which model (if any) performs the best and what conclusions can be drawn from these two analyses. The source code for this analysis is provided in (Marien, 2025).

3. Data analysis

3.1. Garmin activities summary

Data is collected from my Garmin Connect-account (Mariën, 2020), where all track records are stored since 08-2020 (see fig. 3.1).



HARDLOPEN ▾ DOOR KIMI OP JANUARI 28, 2025 @ 4:04 PM

Loon op Zand Hardlopen

Gebeurtenis: -- ▾ Evenementtype: Ongecategoriseerd ▾ Uitrusting: Toevoegen

10.00 km	59:11	5:55 /km	41 m	818
Afstand	Tijd	Gemiddeld tempo	Totale stijging	Calorieën

Figure 3.1. Printscreen Garmin Connect Account

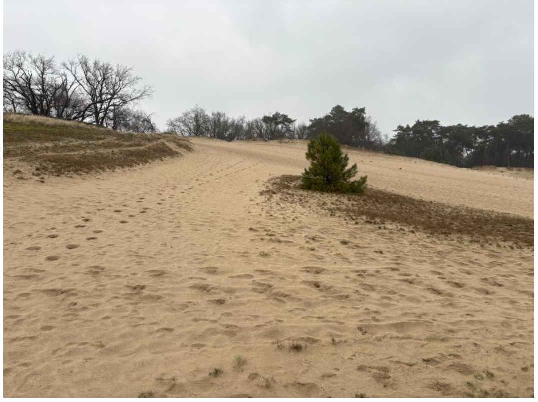
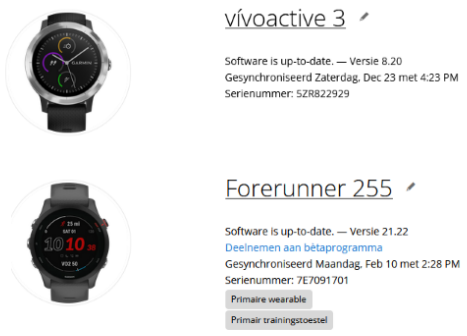


Figure 3.2. Photo of part of the 'Loon op Zand' track with dunes and loose sand

The data is collected using a Vívoactive 3 Smartwatch from



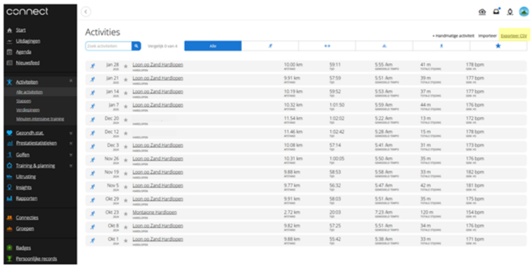
vívoactive 3

Software is up-to-date. — Versie 8.20
Gesynchroniseerd Zaterdag, Dec 23 met 4:23 PM
Serienummer: 5ZR822929

Forerunner 255

Software is up-to-date. — Versie 21.22
Deelnemen aan bètaprogramma
Gesynchroniseerd Maandag, Feb 10 met 2:28 PM
Serienummer: 7E7091701
Primaire wearable
Primair trainingstoestel

Figure 3.3. Examples of the used Garmin Smartwatches



Jan 28	Loon op Zand Hardlopen	10.00 km	59:11	5:55 /km	41 m	178 bpm
Jan 21	Loon op Zand Hardlopen	9.91 km	57:59	5:51 /km	39 m	177 bpm
Jan 14	Loon op Zand Hardlopen	10.19 km	59:52	5:53 /km	37 m	177 bpm
Jan 7	Loon op Zand Hardlopen	10.32 km	1:01:50	5:59 /km	44 m	176 bpm
Dec 20		11.54 km	1:02:02	5:22 /km	13 m	172 bpm
Dec 12		11.46 km	1:02:42	5:28 /km	15 m	178 bpm
Dec 3	Loon op Zand Hardlopen	10.08 km	57:14	5:41 /km	31 m	173 bpm
Nov 26	Loon op Zand Hardlopen	10.31 km	1:00:05	5:50 /km	35 m	176 bpm
Nov 19	Loon op Zand Hardlopen	9.88 km	58:53	5:58 /km	33 m	182 bpm
Nov 5	Loon op Zand Hardlopen	9.77 km	56:32	5:47 /km	42 m	181 bpm
Okt 29	Loon op Zand Hardlopen	9.91 km	58:03	5:51 /km	35 m	175 bpm
Okt 23	Montaione Hardlopen	2.72 km	20:03	7:23 /km	120 m	154 bpm
Okt 8	Loon op Zand Hardlopen	9.82 km	57:25	5:51 /km	34 m	176 bpm
Okt 1	Loon op Zand Hardlopen	9.88 km	55:42	5:38 /km	33 m	171 bpm

Figure 3.4. Export screen Garmin Connect-account

08-2022 until 23-12-2023 and a Forerunner 255 Smartwatch from 01-2024 onwards (fig. 3.3).

Due to the use of 2 different watches, a difference in the accuracy of the measurements might be become an issue but the required accuracy for this analysis is relatively low with kilometers and or multiple meters.

The data used in this analysis is exported using the export button in the online environment which generates the 'Garmin activity description.csv'-file (see fig. 3.4). One manual modification has been made to the file prior to analyzing, which concerns the location 'Kempen'. These tracks start/stop at my current place of residence and are renamed for privacy reasons (grayed out in fig. 3.4).

A complete overview of all columns with description is given in Appendix A. A summary of the recorded tracks is given in figure (3.5).

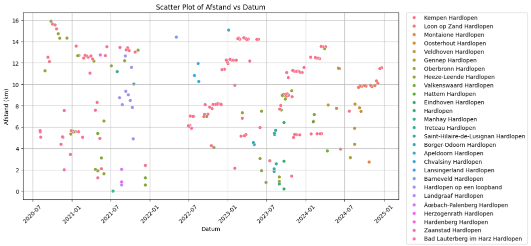


Figure 3.5. All recorded tracks in the period 08-2020 until 12-2024

In figure 3.5, all tracks in the export-file are visualized since 08-2020. The data is consistent with 1 or 2 runs a week, with the exception of the first part of 2022. No tracks were recorded in this period because of a large DIY Home improvement project during which no running took place.

In the scatterplot, color is used as an indicator of the location. Most tracks in 'Kempen' but there are also tracks in Germany, Czech Republic, Italy and France depending on the location of the summer holidays. In the second part of 2024 most tracks are recorded for 'Loon op Zand'. Although there is a big variance, the mean distance is relatively constant in this four year period.

A second observation in figure 3.5 is that the tracks are not equally divided over the different locations. For example, the locations 'Kempen' and 'Loon op Zand' appear more frequent than 'Montaione' or 'Chvalsiny'. This difference will skew the average speed down because less frequent tracks (2 or less) will have a lower average speed than more frequent run tracks. This is mainly because of the way of creating new tracks, especially in new environments. The method usually involves getting audio directions using the Google Maps app on a smartphone. Thus, the first and second time on a new track will involve frequent turn backs or deviations from the planned route. Therefore, since the third time on a new track is usually the first time a consistent speed can be retained, all tracks with 2 or less appearances in the data set will be removed.

Because the data was collected with 2 different smartwatches (see fig. 3.3), not all features are available for the entire dataset. The tracks collected with the VívoActive 3 have less advanced data features than the ones collected with the Forerunner 255. These features include the 'Gemiddelde verticale ratio', 'Gem. verticale oscillatie', 'Gem. grondcontacttijd', 'Gem. GAP', 'Normalized PowerÂ® (NPÂ®)', 'Gem. vermogen', 'Gem. vermogen'. Since these features are not available for any records other than last year's, they will be removed from the data set.

Figure 3.5 highlights one of the major problems with the dataset. Since I don't run tracks that are shorter than 2.5 km, no recorded track should be shorter than 2.5 kilometers. Unfortunately, in the overview there are quite a few recordings below this distance. These indicate or faulty measurements where satellite connection was available for only short window(s) during the track or where the track was manually stopped. In the first case, these are stored as separate tracks in the export-file and are recognizable when there are consecutive tracks within a short time span. Thus tracks within a window of 2.5 hours will merged together. Other tracks shorter than 2.5 km will be removed from the dataset as faulty recordings.

The last major reason for data cleaning is visible in fig.

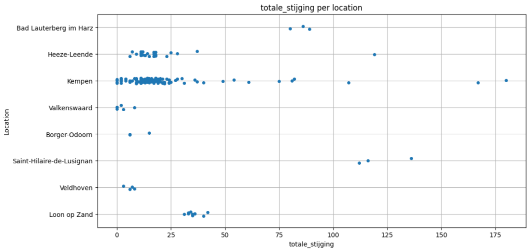


Figure 3.6. Total amount of ascension during each track

3.6. This plot shows the total ascension during a single track. The tracks are grouped per location, and there are seems to be big differences in ascension for certain tracks on the same location. Since the tracks are generally in the same (flat) landscape, values of more than 50 meters can be considered outliers (especially in the Netherlands). There are 2 locations in fig. 3.6 that are not in the Netherlands ('Bad Lauterberg ..' and 'Saint Hilaire ..') which were both in mountainous areas and the meausurements can therefore be regarded as plausible. To correct for these outliers, the median value per location will be used as the new total ascension during a track.

Several other data cleaning actions have been applied to create a good dataset. These are documented in the code (Marien, 2025). The final data set is given in fig. 3.7.

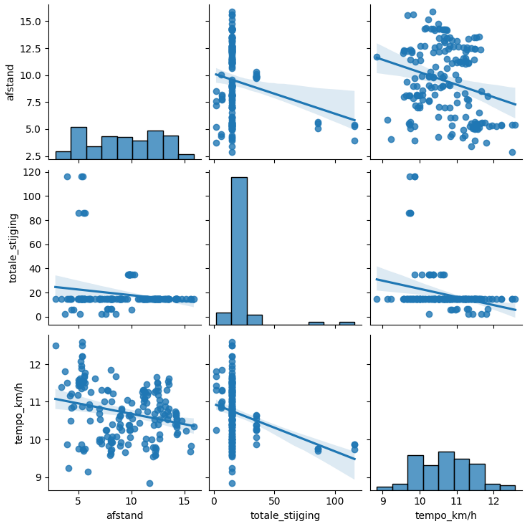


Figure 3.7. Pairplot for the cleaned Garmin Summary File

3.2. Additional features file

For the second regression analysis, a manual csv-file was created with additional features from the Garmin Connect-account that are not exported and visually analysis of the track. For example, in the upper right corner of fig. 3.1 the recorded temperature, windspeed and direction are displayed. These, together with humidy, were not exported to the summary csv-file. The weather data is therefore manually copied from the account to the additional features-csv. Additionally, the track surface was also added to the file. The visualization of atrack in fig. 3.1 shows a recording that had asphalt (bike lanes), loose sands (fig. 3.2) and forest (mountainbike routes) surfaces. The percentage of each of these surfaces is estimated by using the sliders below the google maps visualisation in fig. 3.1. An overview of all columns in the file is given in (Appendix B) and the pair-plot of the numeric columns is given in fig. 3.8.

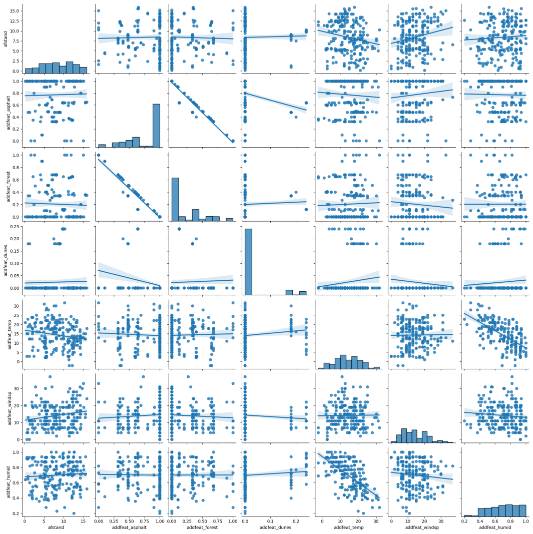


Figure 3.8. Pairplot of the additional summary file

4. Methodology and Implementation

The analysis consists of 2 regression analyses, the first analysis ($location-based\ regression\ analysis$) is based only on the 'Garmin Activities Summary'-file. The data will be preprocessed by cleaning the dataset, removing or adjusting outliers and reformatting data if necessary. From this dataset the target variable $tempokm/h$ will be calculated and stored together with the preprocessed data. This data set will also be used in the second regression analysis.

After preprocessing the data will be normalized and the correlation between the numerical columns and the target variable will be calculated. The correlation matrix will then be used to filter the data set on features with a correlation coefficient of less than 0.19. This means that not all features in the dataset will have a strong correlation with the target variable.

A multivariate regression model will be trained on this reduced dataset. During training, models with a degree up to 3 will be tested and evaluated. To improve the generalization and avoid overfitting, cross-validation (k = 10) and stratification on the target variable will be applied. The models will be run 5 times each time with a different random state.

To assess the performance of the models, both visual plots and calculated metrics are created during testing. The visual assessment of the models will be done by creating 'Actual vs. Predicted' and 'Residual vs. Predicted' plots for all polynomial degrees and random states. To check if bad performances are related to specific instances, the predicted values will be labeled with the instance index in the dataset. The calculated metrics include R^2 and $RMSE$. These will be averaged over all runs to calculate a single performance number.

The second regression analysis ($feature - based$ $regression\ analysis$) will be done in exactly the same way, except that the preprocessed data with the target variable will be merged with the file with additional features. This new merged dataset will also be normalized and filtered on the correlation with the target variable. Regression analysis will be done using the same functions and settings as the first analysis.

5. Evaluation and Results

The results of both regression analyses are presented in this section. For both analyses, the results of the correlation matrix and the final dataset for training of the model will be presented. After that, an example of a visualisation of the best performing model will given and finally a table with the average performance metrics of the models.

5.1. Location based regression analysis

The first 5 rows of the dataset for calculating the correlations between the different features is given in table 5.1.

Date	Location	Dist (km)	Asc (m)	Pace (km/h)
2020-08-04	Bad Lauterberg	5.66	86.0	9.76
2020-08-05	Bad Lauterberg	5.53	86.0	9.70
2020-08-07	Bad Lauterberg	5.05	86.0	9.76
2020-08-27	Heeze-Leende	11.26	15.0	9.89
2020-09-10	Kempen	12.53	15.0	10.41

Table 5.1. Head of data set for location based regression

After cleaning and preprocessing of the dataset, the dataset consists of the target variable, one datetime-column ('datum'), one categorical column ('titel') and two numerical columns '('afstand' and 'totale-stijging'). For the numerical columns, a correlation matrix is calculated and displayed as a heatmap (see fig. 5.1).

Since all features in the dataset have a minimum absolute correlation coefficient with the target variable of 0.25 ($>$ 0.19), no features will be dropped.

Two example visualizations of the tested models are given in fig. 5.2 and 5.3. For all others visualizations, see (Marien, 2025).

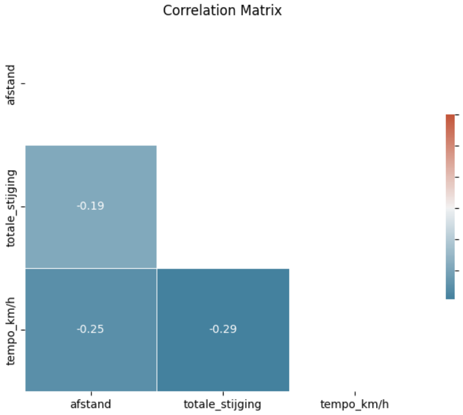


Figure 5.1. Correlation heatmap of the regression analysis dataset

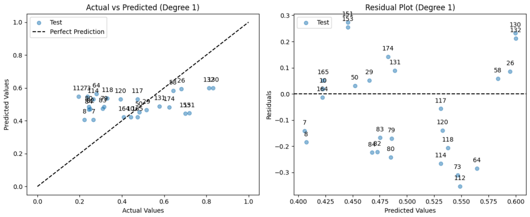


Figure 5.2. Example Pred. vs. Actual for a polynomial d = 1

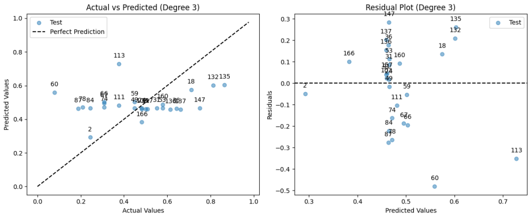


Figure 5.3. Example Pred. vs. Actual for a polynomial d = 3

Looking at the Residual vs. Predicted values-plots (fig. 5.2 and 5.3, right side), there does not seem to be an explicit residual linear relationship in the data left.

Finally, the average results of all runs are given in table 5.2.

	R²			RMSE		
D	Train	Val	Tst	Tr	Val	Tst
1	0.203	0.181	0.002	0.175	0.156	0.192
2	0.267	0.157	0.009	0.168	0.157	0.192
3	0.321	0.184	0.055	0.161	0.154	0.187

Table 5.2. Average Model Performance for Different Polynomials

In the summary table, none of the regression models has a higher test R^2 than 0.1. The polynomial model with d = 3 has the highest Test R^2 with 0.055, and the models with d = 1 and 2 have similar values of 0.002 and 0.009. The models with d = 1 and 2 have the same $RMSE$ of 0.192 which is slightly higher than the 0.187 for the d = 3 model.

5.2. Added features based regression analysis

The first 5 rows of the dataset for calculating the correlations between the different features is given in table 5.3.

Basic Metrics			Surface Type			Weather		
Dist (km)	Asc (m)	Pace (km/h)	Asph	For	Dun	Temp (°C)	Wind (m/s)	Hum (%)
5.66	86.0	9.76	0.90	0.10	0.00	22.2	15	0.38
5.53	86.0	9.70	0.90	0.10	0.00	27.8	9	0.28
5.05	86.0	9.76	0.90	0.10	0.00	31.1	11	0.38
11.26	15.0	9.89	0.64	0.36	0.00	21.1	15	0.53
12.53	15.0	10.41	0.62	0.38	0.00	18.9	11	0.46

Table 5.3. Head of data set for feature based regression

After cleaning and preprocessing of the dataset, the dataset consists of the target variable and nine numerical columns ('Dist', 'Asc', 'addfeat-asphalt', 'addfeat-forest', 'addfeat-dunes', 'addfeat-temp, 'addfeat-windsp', 'addfeat-humid'). For the numerical columns a correlation matrix is calculated and displayed as a heatmap in fig. 5.4.

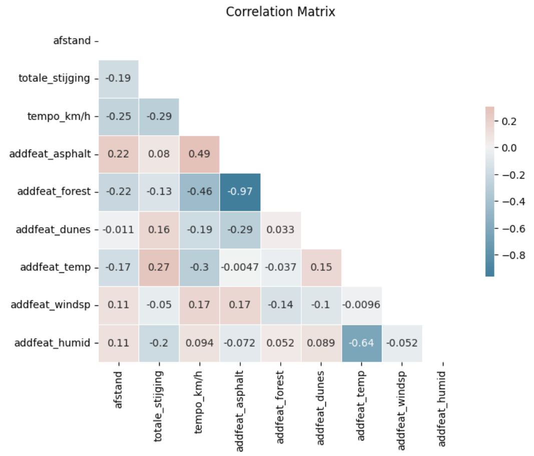


Figure 5.4. Correlation heatmap of the regression analysis datase

There are several features in the dataset that have an absolute correlation coefficient with the target variable of less than the threshold of 0.19 ('addfeat-windsp', 'addfeat-humid'). These variables will be removed. The feature 'addfeat-dunes' is only slightly above the threshold with 0.194 but is kept in the dataset.

Two example visualizations of the tested models are given in fig. 5.5 and 5.6. For all others visualizations, see (Marien, 2025).

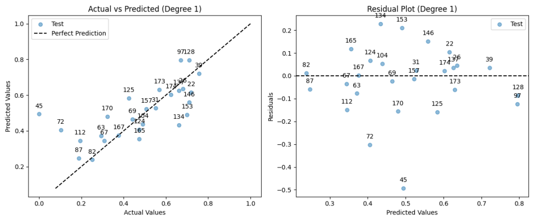


Figure 5.5. Example Pred. vs. Actual for a polynomial d = 1

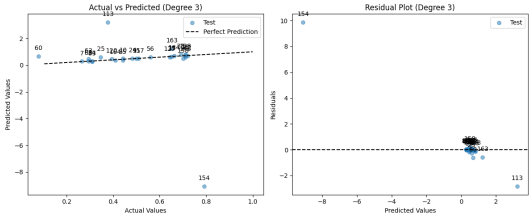


Figure 5.6. Example Pred. vs. Actual for a polynomial d = 3

Looking at the Residual vs. Predicted values-plots (fig. 5.5 and 5.6, right side), especially the d = 3 model predicts large residuals.

Finally, the average results of all runs are given in table 5.4.

	R^2			RMSE		
D	Train	Val	Tst	Tr	Val	Tst
1	0.652	0.503	0.552	0.115	0.119	0.128
2	0.713	0.489	0.572	0.105	0.121	0.124
3	0.789	-2768.0	-31.34	-1.4×10^9	4.589	0.826

Table 5.4. Average Model Performance for Different Polynomials

In the summary table, none of the regression models has a higher test R^2 than 0.6. The polynomial model with d = 2 has the highest Test R^2 with 0.572, followed by the d = 1 model with 0.552. The model with d = 3 has a large negative R^2 value of -31.34. The best performing model based on $RMSE$ is also model with d = 2, with a $RMSE$-value of 0.124 (compared to 0.128 for the d = 1 model). The d = 3 model has the highest $RMSE$-value of 0.826.

6. Conclusions and Discussion

During this data analysis project several interesting findings were discovered. Firstly, a (multivariate) linear model does not seem to be the correct model to predict the expected average speed. Even with the additional features such as surface features and temperature, the R^2 improved only to 0.58 and the $RMSE$-value to 0.124 (it is important to note that $RMSE$ is calculated on the normalized data and thus do not correspond directly to a value in km/h).

Secondly, it is interesting to note that the model with polynomial d = 1 performs almost as good as the d = 2 model. If a linear model is going to be used, the preference would be to use the simpler d = 1 model. The model with d = 3 performs the worst with a large negative R^2. This means that the model is too complex, leading to overfitting the data, large weights and performing worse than predicting the average values.

Thirdly, for both regression models the distance ('afstand') is not the variable with the highest correlation with average speed. In the first model the ascension has the highest correlation and in the second analysis, asphalt had the highest correlation. It is surprising that the variable 'dunes' has a correlation of only 0.194, given the experiences of the author on these surfaces (fig. 3.2). Looking at figure 3.1, the upper part of the photo shows a Google Map with a high dune with very loose sand, which corresponds to a big dip in speed in the lower part of the photo (and where fig. 3.2 was taken). The low correlation to average speed might be caused by the relative few tracks with dunes in the data set.

Future improvements of this analysis could be made by looking at non-linear models such as Support Vector Machines for Regression (SVR) or Regression Trees/Random Forests. A second improvement would be to look for more fine grained data such as the length and steepness of individual slopes in a specific track instead of the total ascension. A final improvement might be to collect more data with the Forerunner 255 since variables like 'Gem. GAP' are calculated on this data. Since this is a corrected metric for hills etc, it might have a better correlation with the average speed (Appendix A).

References

Acosta, G. M. T. Using python and machine learning to analyze my running. part 1. https://gmauricio-toledo.medium.com/using-python-and-machine-learning-to-analyze-my-running-8c42c34744a8, 4-4-2021. Accessed: 2024-12-18.

Cam, R. where is part 2 please ? https://medium.com/@roncam082/where-is-part-2-please-4eee14916d49, 20-3-2022. Accessed: 2024-12-18.

Leuven, P. K. Topsport abc begeleidt deelnemers aan canvas-programma 'marathon'. https://nieuws.kuleuven.be/nl/2012_en_vroeger/0607/02/topsport-abc-begeleidt-deelnemers-aan-canvas-programma-2018marathon2019, 10-10-2006. Accessed: 2024-12-18.

Marien, M. Source code for ml assignment, 2025. URL https://github.com/KiMideVliet/Machine-learning.git. Accessed: 2025-02-11.

Mariën, M. Activities. https://connect.garmin.com/modern/activities, 2020. Accessed: 2024-12-20.

A. Overview of columns in the Garmin export file

Index	Column	Non-Null Count	Dtype	Description of feature
1	Activiteittype	226 non-null	object	Sport of activity [-]
2	Datum	226 non-null	object	date and time of activity [-]
3	Favoriet	226 non-null	bool	Activity marked as favorite [-]
4	Titel	226 non-null	object	Garmin added location [-]
5	Afstand	226 non-null	float64	Distance of track [km]
6	Calorieën	226 non-null	object	Burned calories during activity [calories]
7	Tijd	226 non-null	object	Duration of activity [H:M:S]
8	Gem. HS	226 non-null	object	Average heart rate during activity [pcs/min]
9	Max. HS	226 non-null	object	Max heart rate during activity [pcs/min]
10	Training effect aeroob	226 non-null	object	Aeroob trainingseffect [-]
11	Gem. loopcadans	226 non-null	object	Average running cadence [steps/min]
12	Max. loopcadans	226 non-null	object	Max. running cadence [steps/min]
13	Gemiddeld tempo	226 non-null	object	Average pace [min/km]
14	Beste tempo	226 non-null	object	Best pace [min/km]
15	Totale stijging	226 non-null	object	Total ascension [m]
16	Totale daling	226 non-null	object	Total descending [m]
17	Gem. staplengte	226 non-null	object	Average stepsize [m]
18	Gemiddelde verticale ratio	226 non-null	object	Average vertical ratio [-]
19	Gem. verticale oscillatie	226 non-null	object	Average vertical oscillation [m]
20	Gem. grondcontacttijd	226 non-null	object	Average ground contact time [ms]
21	Gem. GAP	226 non-null	object	Grade-Adjusted Pace [min/km]
22	Normalized Power® (NP®)	226 non-null	object	Cycling-specific power adjustment [watt]
23	Training Stress Score®	226 non-null	float64	Quantification of overall training load [-]
24	Gem. vermogen	226 non-null	object	Average power [watt]
25	Max. vermogen	226 non-null	object	Max power [watt]
26	Decompressie	226 non-null	object	Decompression [Y/N]
27	Beste rondetijd	226 non-null	object	Best time per km [H:M:S]
28	Aantal ronden	226 non-null	int64	Number of km's [pcs]
29	Tijd bewogen	226 non-null	object	Time moved during activity [H:M:S]
30	Verstreken tijd	226 non-null	object	Total Time of activity [H:M:S]
31	Minimum hoogte	226 non-null	object	Min. height during activity [m]
32	Maximum hoogte	226 non-null	object	Max. height during activity [m]

B. Overview of columns in the additional features file

Index	Column	Non-Null Count	Dtype	Description of feature
1	datum	216 non-null	object	Date of activity [-]
2	plaats	216 non-null	object	Location of activity [-]
3	afstand	216 non-null	float64	Distance of track [km]
4	addfeat asphalt	216 non-null	float64	Amount of asphalt as track surface [percentage]
5	addfeat forest	216 non-null	float64	Amount of forest trails as track surface [percentage]
6	addfeat dunes	216 non-null	float64	Amount of dunes as track surface [percentage]
7	addfeat temp	216 non-null	float64	Temperature [Celcius]
8	addfeat sky	216 non-null	object	Description of weather [-]
9	addfeat windsp	216 non-null	int64	Windspeed [km/u]
10	addfeat winddir	226 non-null	object	Wind direction [-]
11	addfeat humid	216 non-null	float64	Humidity [percentage]