Guillaume Wägli
1700 Fribourg
Guillaume.waegli@gmail.com

Data Science Project

Run pace predictor

Conceptual Design Report

5 October 2025

Abstract

Sports watches today collect extensive data from training sessions, enabling athletes to monitor performance and even receive predictive race times. However, a lot of existing predictors are limited to standard distances and fail to account for other factors like elevation or individual variability. This project aims to address these shortcomings by developing a mini-application capable of estimating personalized race times using past activity data combined with specific race parameters. The solution leverages supervised machine learning techniques, employing regression models such as Linear Regression, Random Forest, TensorFlow, and XGBoost to capture relationships between distance, speed, elevation, and other recorded variables.

The initial dataset consists of over 170 sessions recorded by a Garmin Descent watch, with more than 100 characteristics per run. Data preprocessing and analysis will be performed in Python using libraries such as Pandas, NumPy, and scikit-learn, followed by model training and validation. A graphical interface will allow users to input race details and visualize predictions alongside personalized statistics. In a second phase, the project will explore extending predictions to datasets from multiple athletes, raising challenges of generalizability and data consent.

Table of Contents

1 Project Objectives

Current sports watches record a wide range of data during activities. They allow athletes to track their progress. Some watches also use this data to offer personalized training sessions aimed at improving athletes' abilities. Based on all this data, some watches are even capable of predicting potential race times in advance.

Figure 1: Garmin race predictor

Unfortunately, the estimated times are only for "standard" distances and do not take into account elevation changes or other factors that could affect race times.

The first objective of this work will therefore be to create a mini application that, based on training data as well as the parameters of an upcoming race, can estimate a potential run time.

In a second phase, if the first objective is fulfilled, others data set will be used to predict the race pace from other runners. The calculations will be then adapted to fit various athletes.

2 Methods

To best estimate race times based on past data, supervised machine learning will need to be used.

Information such as distance, speed, and elevation will be important for performing regressions. However, the watch records many other parameters, so special attention must be paid to all available data.

Initially, the code will be developed in Python using Google Colab. The following libraries/modules will be required:

- **Pandas.DataFrame** (to structure data in a 2-dimensional table)
- **NumPy** (for numerical calculations)
- **Matplotlib** (for creating graphs)
- **scikit-learn** (for preparing data, normalization, encoding, train/test splitting)
- **LinearRegression** (for regression)

- **RandomForestRegressor** (for regression)
- **TensorFlow** (open-source machine learning platform)
- **XGBoost** (for building predictive models by combining multiple "weak" decision trees.)

The next goal is to add an interface to allow users to input parameters, click a button, and obtain a result. For this, it is planned to use **Tkinter** or a similar interface.

3 Data

The initial input is a JSON file containing all the information from every activity recorded by the watch. To access the data, an extraction from the Garmin.com website is required, which take at least 48 hours. Once the base dataset is loaded, it will be considered to allow manual uploading of activities from the app, without needing to extract everything each time.

The data used are recorded from a Garmin Descent watch. The earliest data dates back to December 2022, and the most recent data currently available are from September 23, 2025. The dataset consists of 153 running sessions and 19 trail sessions.

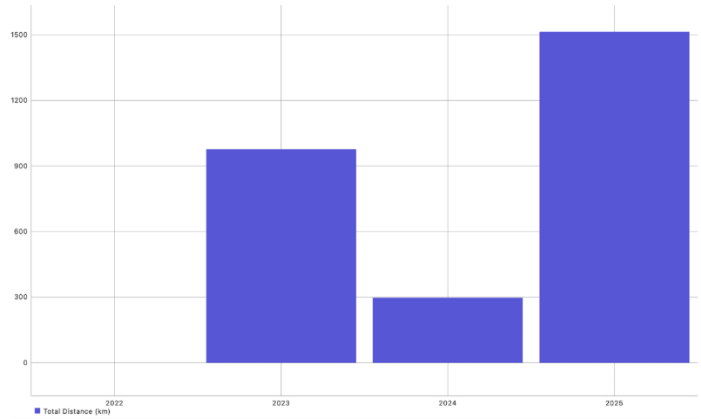

Figure 2: Distance covered in three years, spread over 153 sessions

Each line represents an activity (run). However, the watch records numerous characteristics that form columns. There are therefore 110 characteristics for each line. Not all of them will be necessary, but after an initial analysis, it appears that at least 57 of them could be useful.

As the data belongs to the author, there are no security issues associated with its use for the first time of the project. In the second part, the author must obtain consent from individuals for the use of their data.

4 Metadata

For the time estimation to work, it is essential that all loaded data belong to the same athlete. This can, in principle, be verified by filtering the value of **"userProfileId"**.

Since this is a personal estimate, external users are not expected to have access to the dataset.

However, anyone with a Garmin sports watch should be able to extract their own data and calculate their own estimate. To check the accuracy of the results, a comparison with the race predictor (Figure 1) can then be made.

5 Data Quality

Since the data is recorded by the watch during activities, there should, in principle, be no missing values in the data set. The initial data check shows no missing values.

However, data quality issues could arise if the watch records data incorrectly or inaccurately. For example, if the GPS is poorly calibrated, distances or elevation changes could be incorrect, which would skew the calculations.

To ensure the watch is functioning properly, occasional comparisons can be made during official races where time and distance are measured by external infrastructures. If necessary, the watch should be adjusted or replaced.

6 Data Flow

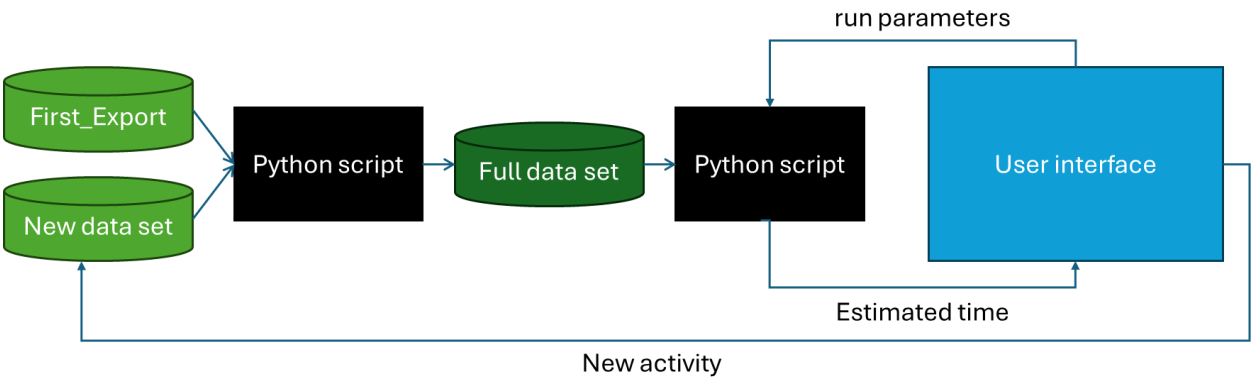

Figure 3: data flow

How It Works:

The user interface has several functions. The first is to allow new activities to be loaded so that they are considered when calculating the race time. The interface also allows users to input race

parameters so that the script can then calculate the estimated required time. The time calculated by the script is also displayed through the user interface.

The user interface will also provide the user with various statistics, such as distances covered, average speed, and the evolution of metrics over recent activities.

7 Data Model

The conceptual model is to develop a tool that can reliably estimate the time required to complete a race based on past training.

At the logical level, it is already clear that data such as speed, distance, and elevation from past activities will be essential information. However, the available data is extensive, and part of the work will be to determine whether it can contribute to improving predictions. Ultimately, this data will be used in a regression model.

At the physical level, calculations can be performed with a basic infrastructure. However, a “complete” dataset, including both existing activities and newly recorded activities, will need to be created. A document stored on Google Drive is planned for this purpose.

8 Documentation

Before carrying out the project, a review of the different methods for estimating race times will need to be conducted. This research will be made available to interested parties and can serve as a basis for calculations.

Regarding the project documentation itself, it is planned to describe the code as thoroughly as possible to maintain a clear understanding of the workflow and to ensure reproducibility.

9 Risks

At this stage of the project, three main risks have been identified:

1. **Lack of data:** To calculate estimates for all possible scenarios, a sufficient volume of varied runs must be available. Data that is too homogeneous could make it difficult to estimate very different types of runs.
2. **Old data:** Since physical condition evolves over time, particularly depending on activity, older data might not reflect current fitness and could lead to inaccurate estimates.
3. **Risk of non-reproducibility:** Currently, we only have measurement data for a single subject. There is therefore a risk of developing a solution that is tailored to this individual but does not

work for predicting times for other individuals. Ideally, predictions should be tested with additional subjects.

Erroneous estimates could also occur if a user falsifies their activity data. It is therefore necessary to make individuals aware that the quality of the predictions also depends on the accuracy of the information they provide.

10 Preliminary Studies

While this work is relevant to the development of a solution and the application of coding and statistical skills, a great deal of research has already been conducted in this area, and numerous online applications/solutions already exist for calculating time.

Before embarking on the development of the solution, it seems important to consider what has already been done and what is commonly considered.

Formule de Riegel (exponentielle de performance)[1] :

$$T_2 = T_1 \times \left(\frac{D_2}{D_1}\right)^k$$

Figure 4: Riegel Formula

Riegel's formula is a basic formula in which T_1 represents the known time, D_1 the corresponding distance, T_2 the target distance, and k the empirical coefficient, generally 1.06.

Maximal aerobic speed (MAS)[2] :

MAS or VMA is the speed at which an individual reaches their maximum oxygen consumption. VMA is a key performance indicator for endurance training, particularly for interval training, and is calculated using specific physical tests like the cooper test.

Knowing the MAS allowed then to determine a running time based on the distance foreseen:

[1] (Dash, 2024)
[2] (Walker, 2025)

VMA	TEMPS SUR 10KM (85-90% VMA)	SEMI-MARATHON (80-85% VMA)	MARATHON (75-80% VMA)	Allure max footing (70%VMA)
8 km/h	1h23m20s à 1h28m10s	3h6m0s à 3h17m37s	6h35m14s à 7h1m57s	10m42s au km
9 km/h	1h14m0s à 1h18m20s	2h45m16s à 2h55m49s	5h51m38s à 6h14m50s	9m31s au km
10 km/h	1h6m40s à 1h10m30s	2h28m44s à 2h38m14s	5h16m28s à 5h37m34s	8m34s au km
11 km/h	1h4m10s à 1h30s	2h15m22s à 2h23m49s	5h6m37s à 4h47m38s	7m47s au km
12 km/h	58m40s à 55m30s	2h3m46s à 2h11m51s à	4h41m18s à 4h23m43s	7m8s au km
13 km/h	51m10s à 54m10s	1h54m17s à 2h1m40s	4h3m19s à 4h19m30s	6m35s au km
14 km/h	47m30s à 50m20s	1h46m11s à 1h52m52s	3h45m45s à 4h31s	6m7s au km
15 km/h	44m20s à 47m0s	1h39m9s à 1h45m29s	3h30m59s à 3h45m2s	5m42s au km
16 km/h	41m40s à 44m0s	1h32m50s à 1h38m48s	3h17m37s à 3h30m59s	5m21s au km
17 km/h	39m10s à 41m30s	1h27m33s à 1h32m50s	3h5m39s à 3h18m19s	5m2s au km
18 km/h	37m0s à 39m10s	1h22m38s à 1h27m54s	2h55m49s à 3h7m4s	4m45s au km
19 km/h	35m0s à 37m0s	1h18m4s à 1h22m59s	2h45m58s à 2h57m13s	4m30s au km
20 km/h	33m20s à 35m10s	1h14m11s à 1h19m7s	2h38m14s à 2h48m47s	4m17s au km
21 km/h	31m40s à 33m30s	1h10m40s à 1h15m15s	2h30m30s à 2h40m20s	4m4s au km

Figure 5:table estimating running time based on VMA[3]

Other methods exist, notably those based on heart rate or using variations of the Rigel formula[4] or based on similar methods shown before. However, it appears that regression tests are always necessary to estimate a running time with greater accuracy.

11 Conclusions

There are already many ways to predict race times. Even Garmin offers this in its “basic” functions. However, the ability to incorporate variables such as elevation or specific distances makes developing this tool interesting and potentially useful to many people.

There is little doubt about the feasibility of this work. However, it is not certain that the results will be reliable or truly usable. The solution will likely need to be tested and adapted several times.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu

[3] (Decathlon Blog, n.d.)
[4] (Dash, 2024)

speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 04.10.2025 Signature(s): G. Wägli

Bibliography

Dash, S. (9. October 2024). *Win Your Race Goal: A Generalized Approach to Prediction of Running Performance.* Von National Library of Medicine: https://pmc.ncbi.nlm.nih.gov/articles/PMC11495242/ abgerufen

Decathlon Blog. (n.d.). Retrieved from Qu’est-ce que la VMA en sport ?: https://conseilsport.decathlon.fr/quest-ce-que-la-vma-en-sport

Walker, O. (25. 03 2025). *Science for sport.* Von Maximal Aerobic Speed (MAS): https://www.scienceforsport.com/maximal-aerobic-speed-mas/?srsltid=AfmBOoodX6xD8kpoI6pVM-at8lQx1OR5e7dbdfS-qlBFZtTMGUe3FA6X abgerufen

List of figures