

Universiteit Leiden

Master Computer Science

Explainable variables for Fitness Monitoring in Professional Cycling.

Name:	F.J. Haasnoot
Student ID:	1716689
Date:	12-01-2026
Specialisation:	Data Science
1st supervisor:	Dr. A.J. Knobbe
2nd supervisor:	A.W. de Leeuw

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Contents

1 Introduction **3**

2 Preliminaries **6**
2.1 Related work 6

3 Methods **9**
3.1 Target definition 9
3.2 Searching for other predictors 10
3.3 Model 12

4 Experimental setup **13**
4.1 Data 13
4.2 Feature groups 15
4.3 Model setup 15

5 Results **18**
5.1 Clustering and regression 18
5.2 Classification 25

6 Discussion **31**

7 Appendix **33**

Abstract

In professional cycling, it is very important to know the current state of a rider. The power heart-rate model is a model for continuous fitness monitoring to asses such a state. However, this model only uses pedal power as a sole predictor, while the heart rate of a cyclist is influenced by a variety of factors. In this report, I conducted research on improving this power heart rate model by looking at additional explanatory variables. The goal was to discover what kind of variables affect heart rate in order to improve the performance of the model. After doing multiple regression and classification experiments, factors as elapsed time training sessions, distance ridden and traversed altitude meters had a big impact on the power heart rate model. To improve performance of the power heart rate model, these factors should be included.

1 Introduction

For professional cyclists to be at their best during races, they need to be in optimal physical condition. To achieve such a condition, the balance between training stress and recovery needs to be optimized (Kellmann, 2010). Each cyclist has its own plan with different training loads to reach such a fitness level. These plans are highly adaptive because training with a too small or too large training load increases the risk of injury (Bourdon et al., 2017). Monitoring is an important step to define the balanced load and this can be done in various ways. With newer technologies such as power meters being widely available, cycling is an ideal sport for continuously monitoring athlete performance and acting accordingly. To assess the level of a cyclist, all kinds of training load need to be considered especially at the higher levels, since these are levers required during a race. One metric to judge the level of a cyclist is Functional Threshold Power (FTP). This is the maximum average power output sustained over a period of 1 hour. Tests were developed, for example the 20 minute FTP test, which was deemed representative of a highly trained athlete to assess its FTP(McGrath et al., 2019).

In contrast to the FTP test that measures power as a performance indicator, heart rate is also used as an indicator for three important metrics, namely VO_2max, lactate threshold 1 (LT1) and lactate threshold 2 (LT2) (Joyner & Coyle, 2008). VO_2max refers to the maximum amount of oxygen that the athlete can absorb and use during exercise. LT1 and LT2 (both

expressed in heart rate), respectively, also known as aerobic and anaerobic thresholds, are thresholds on how the body manages lactate. LT1 is the moment where the body starts to release lactate in the blood above a certain baseline, and LT2 is the moment where lactate is largely starting to build up. These metrics can be assessed in a lab but require the athlete to exert maximum effort, making them demanding for athletes. Such efforts can be a disruption of the ongoing training plan and set the athlete back in their progress, and due to the difficult nature of these tests, athletes generally dislike them(Fry et al., 1991). Continuous fitness monitoring of all training sessions and race's would be a solution to this problem because it would not require any testing to assess the cyclists' level. An athlete monitoring approach was developed by De Leeuw et al. which was based on a daily power heart rate model (PHR-model) that couples external training load with internal training load (de Leeuw et al., 2023) External load here is the result of training of an athlete, e.g., the power output of a cyclist, during their sessions, and internal load is how the body of that athlete reacts to this external load, for example, heart rate. This PHR-model describes the relationship between heart rate (bpm) and power output (wattage). Looking at this relationship, a performance team could see if a cyclist is on track by comparing different rides on wattage output and the corresponding heart rate. If the heart rate was higher at a specific wattage than in other sessions, this could, for example, be a sign of overtraining. With data from training sessions and the PHR-model, you could approximate the VO_2max, LT1 and LT2 values. Due to the PHR-model, you could assess the level of the riders after each session and thus eliminate the need for frequent testing.

The results of the power PHR-model show that there are anomalies in the data which the model could not explain. The model only uses convolved pedal power as a predictor of heart rate. This is the main relationship, since power is a direct result of the body working. However, it is likely that other factors also influence how the body reacts during exercise, which means that there are other predictors available for heart rate in addition to the convolved pedal power. In this research, I will try to improve the model by looking for other explanatory variables for the PHR-model. One hypothesis is that the geographic location and especially mountainous area's influence the heart rate because these area's contain climbs of the highest categories in terms of steepness and length. In addition to geographic variables, other predictors are also taken into account such as length of the training session or at which cadence the rider is cycling. By performing a regression and classification

analysis, new predictors can possibly be found to add to the model to increase performance. In conclusion of the above, I will do the following experiments:

- Cluster analysis will be performed to extract the primary geographical locations visited within the dataset. Based on these clusters, distance features will be created by computing the distance between each cluster and the actual position.
- A regression task will be conducted to assess the contribution of all potential predictors and to verify that the model fully captures their explanatory power.
- A classification task will be carried out to explore whether additional variables can serve as predictors and thereby improve the model.

First, I will discuss some related work on this subject. I will explain how convolution of the pedal power of the PHR-model works and how a relationship is modeled between the convolved power and heart rate. Then, I will elaborate on my own methods and experimental setup and discuss the results. I will end with my conclusion and discussion.

2 Preliminaries

This paper builds on the work of De Leeuw et al. and the subsequent research by Damstra et al. In the following, I provide a concise overview of their approach to convolving pedal-power data and the kernels used in this process, concluding with the power heart-rate model.

2.1 Related work

To accurately model the heart-rate response during exercise, it is necessary to apply an aggregation step that captures the influence of past efforts. Heart rate does not depend solely on power output, but also on accumulated load and recovery from previous exertion. By convoluting the pedal power of a certain window and sum the resulting values, it is possible to define a direct relationship between heart rate and pedal power (de Leeuw et al., 2023). Convolution is a mathematical operation that combines two functions to produce a third one (Hirschman & Widder, 2012). In this context, it involves combining a sequence of pedal-power values, taken over a specified window, with a kernel function to generate a transformed sequence. For a session s of length at time t, the convolved pedal power is defined as:

$$P^s_{\text{conv}}(t) = \frac{\sum_{i=0}^{w} h(-i)\, p^s(t-i)}{\sum_{i=0}^{w} h(-i)} := \sum_{i=0}^{w} \tilde{h}(-i)\, p^s(t-i)$$

– Here w is the length of the convolution window, $p^s = \{p^s(0), p^s(1), \ldots, p^s(l)\}$ is the pedal power attribute, and $h(x)$ is the kernel. The kernel weights are normalized so that the sum of the weights is 1. This is to obtain a convolved pedal power on the same scale as the pedal power of the data set.

To obtain a good convolution of pedal power, different kernels were suggested. One approach was to set the convolution window at -180 seconds and construct a kernel where each second and their respective pedal power represent a predictor with heart rate at $t = 0$ as the target variable. For this to work, the convolution window needs a variable for each second. This means that the first 180 seconds were removed from each session since they had no 180 previous records. In addition, interpolation was performed to ensure that each window is completely filled. If too many data points were missing and interpolation was not possible, the data was removed until the next proper part of the data set. Subsequently, this new part also had to have the first 180 seconds removed to work.

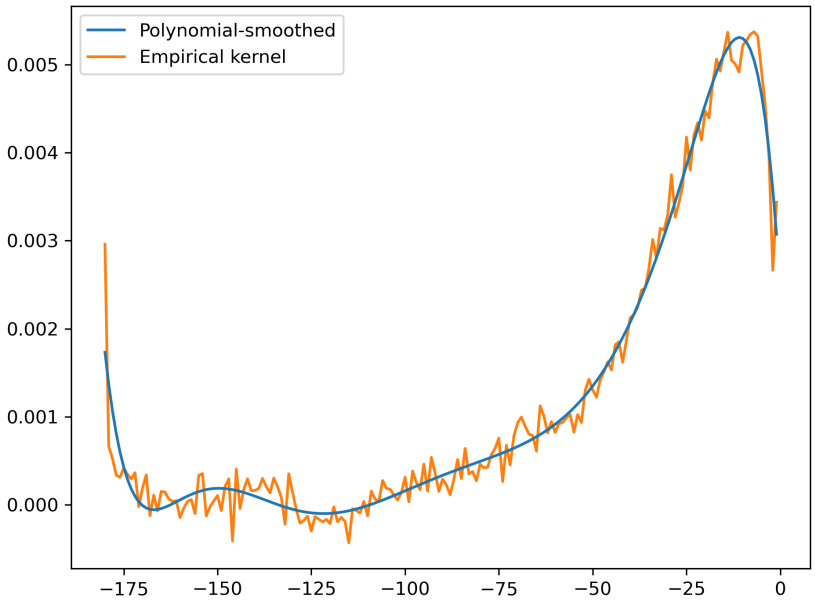


Figure 1: The results of the multiple linear regression and the 10th polynomial smoothing of a random training session.

Then a multiple linear regression (MLP) was performed for each session:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

– Here y denotes the dependent variable, x_j the independent variables, β_j the regression coefficients and e the model error. The MLP gave coefficients for each second and their pedal power value. Later, these coefficients were used as weights to multiply them by the actual pedal power to retrieve the convolved pedal power. In the research of Damstra et al. they compared different kernels on their performance. The kernel where MLP was used to get the coefficients of the function and then a smoothing by a 10th order polynomial (polynomial kernel) which you can see in Figure 1, gave the best results compared to other kernels. For this reason, in this research the polynomial kernel will be used.

After convolution, there is a direct relationship between the convolved pedal power and heart rate. To model the relationship where pedal power functions as an independent variable and heart rate as a dependent variable, De Leeuw et al proposed an S-shaped curve. This is typically calculated as

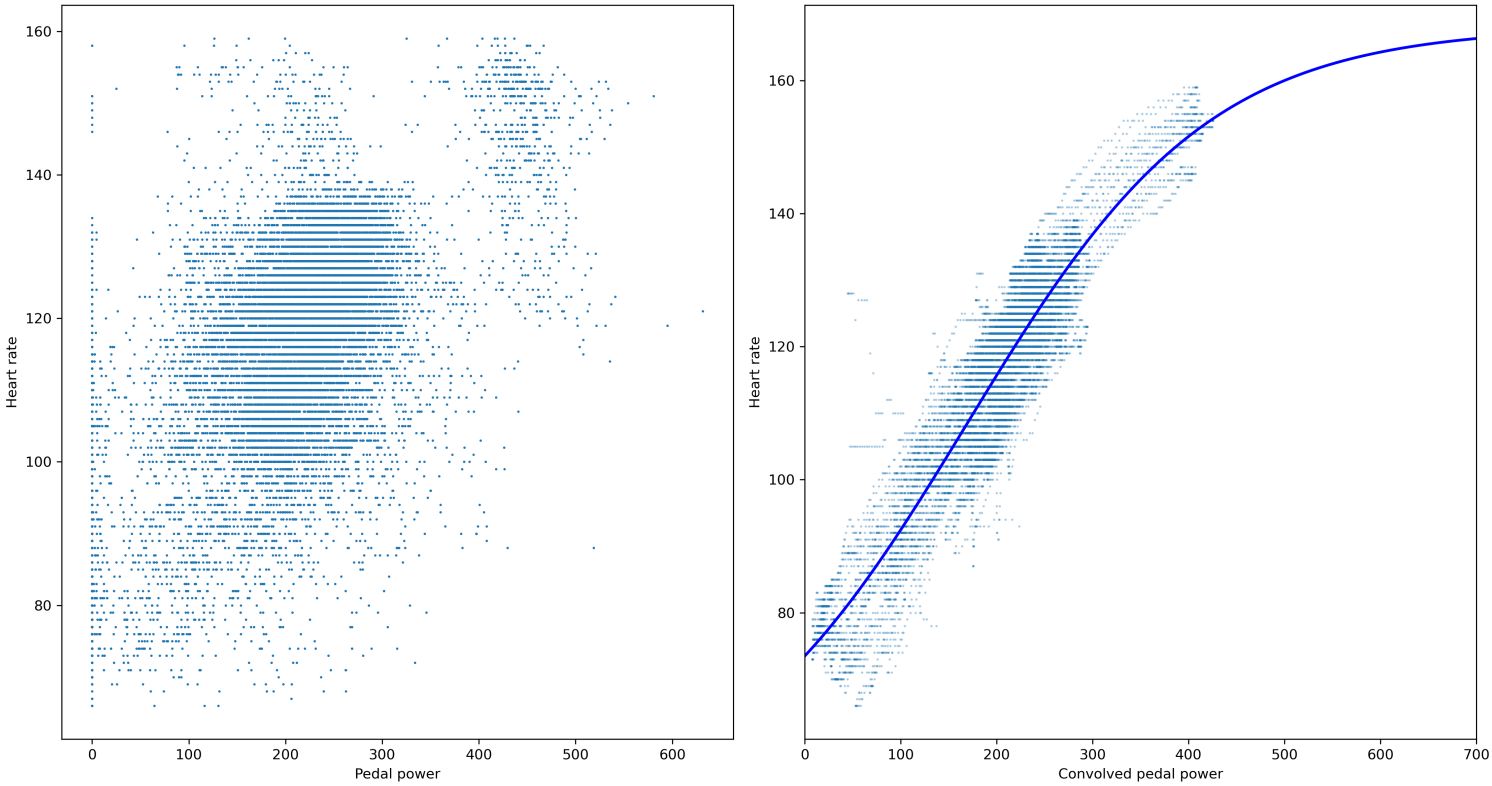


Figure 2: The PHR-model of a random training session. Onto the left heart rate as a function of pedal power and onto right heart rate as function of convolved pedal power with a sigmoid curve fitted.

a sigmoid function:

$$f(x) = A + \frac{L - A}{1 + e^{-k(x - x_0)}}$$

– where L denotes the upper asymptote, A denotes the lower asymptote, k the steepness of the curve, and $x0$ the x value of the middle point. The parameters of these functions are optimized for each session to fit the most appropriate curve.

With the newly created function, it is possible to compare the performance of different cycling training sessions.

3 Methods

As discussed in chapter 2, the PHR model defines a relationship between convolved pedal power and heart rate. After this, a sigmoid curve is fitted, where certain information can be extracted subsequently, like lactate threshold, for example. If the model were perfect, all the data points would lie exactly on the sigmoid curve, which is not the case as seen in Figure 2. Although several factors may explain these deviations, the most plausible explanation is that the model lacks information to explain the observed variance in heart rate. To evaluate how additional variables influence the predictive performance of the PHR-model, three experiments were conducted. These experiments aim to examine whether factors beyond power output contribute meaningful explanatory information and whether incorporating such variables can enhance the accuracy of the model. Together, they provide a framework for evaluating the degree to which the current modeling approach captures the underlying dynamics present in cycling data. In the following sections, I will elaborate on each experiment in detail.

3.1 Target definition

In order to do the supervised machine learning experiments, the target variable needs to be defined. As we want to discover which other variables impact the PHR-model, it is logical to use the error of this model output as a target for the experiments of this research. These residuals, the vertical differences between the curve and the actual data points which can be seen in Figure 2, will be the target of these experiments. If the experiments showed some correlation between the predictor variable and the residuals, this could mean that these predictor variables could enhance the PHR model by incorporating these features as predictors.

For regression experiments, these raw residuals can be used as a target predictor but not for classification. In the original research, it was noticeable that there were sequences of successive heart rate data points that were far above or below the curve. The regression experiment is important because it possibly shows if other variables influence the PHR-model. However, this model is also trained on smaller deviations of, for example, 1 or 2 heart rate points. These deviations can be explained by different variables than a deviation of 10 heart rate points. That is why a classification experiment was conducted to see if this would give different results compared to the regression

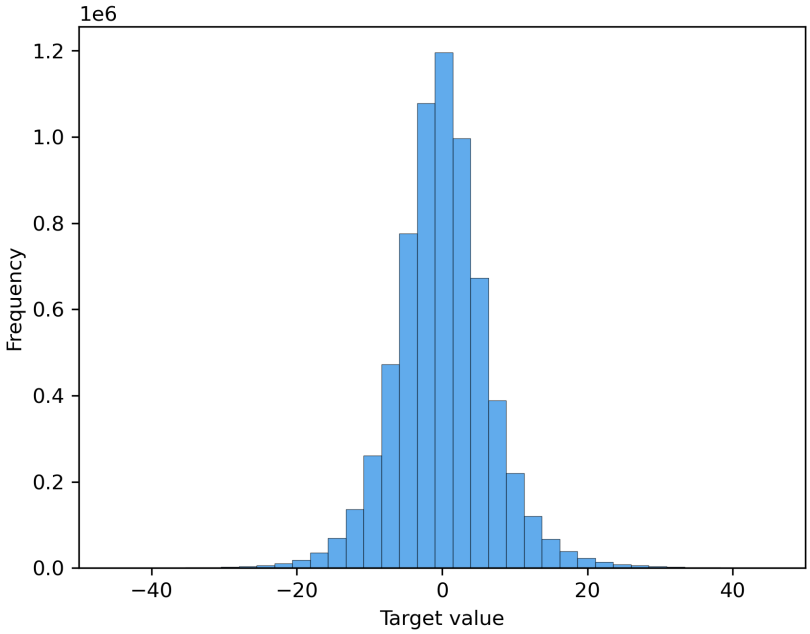


Figure 3: Distribution of the target variable. Targets are residuals measured in heart rate (BPM) from the PHR-model. Outliers above 50 and below -50 are not shown here but will be taken into account for the experiments.

experiment. For the classification experiment, a binary target variable was constructed. This target was defined by thresholding the response variable at an absolute value of 10 and by adding a temporal aspect. Specifically, a sample y at time t was assigned to the target class if

$$|y_{t-1}| > 10 \ \wedge \ |y_t| > 10 \ \wedge \ |y_{t+1}| > 10,$$

and to the non-target class otherwise. Boundary samples were excluded from target construction due to the absence of adjacent samples. After constructing the binary target variable, 6.291.945 cases were considered non-target and 343.596 as target.

3.2 Searching for other predictors

Additional information about training sessions can be found in numerous ways. Since one of the original goals of the PHR-model research was to

assess the level of a cyclist not only by testing but also monitoring training rides, our data set contains only data in an uncontrolled outdoor setting. This means that there are many external factors that could possibly have an effect on the rider. One type could be the weather variables. For example, higher temperatures are known to induce a higher oxygen demand, resulting in a higher heart rate (Galloway & Maughan, 1997). Other research showed that the combination of increasing temperature and humidity caused a higher oxygen demand with an increase in heart rate of 3-6%. So, at an average heart rate of 150, this would mean an increased heart rate of 4.5 to 9 BPM (Nazaret et al., 2023). These variables have a physiological effect on how the body performs under different circumstances, but certain variables may also have an effect on the mental state of the rider. Mental stressors are known to increase heart rate (Taelman et al., 2009). In cycling, it could be scary to descend steep roads, especially during training sessions, when normal cars are also on the road. The resulting stress could possibly increase the heart rate without the need for extra power on the pedals. Then it is important to know that there is a phenomenon called cardiovasculair drift in which the heart rate increases during prolonged exercises when the intensity remains the same (Coyle & Gonzalez-Alonso, 2001). It could be that the elapsed riding time or distance that has been covered during the ride increase the residuals of the model.

Geographical location could also influence the heart rate of a cyclist. If a road has a certain gradient, cyclists need to deliver more power to achieve the same speed compared to a road with zero gradient, which is also the case when comparing situations with or without a headwind. Since the model only describes the relationship between power and heart rate and speed is no variable, it should not make a difference for the cyclist's heart rate if the cyclist is on a steep road or a flat road with a lot of headwind, as long as they deliver the same power output. You could say that the geographical location should not have an impact on the model as long as the required power output is the same.

However, there are climbs infamous for their length and gradients. One of the Mont Ventoux climbs for example has a length of 20.8 kilometers and an average gradient of 7.7% with the steepest 100 meters of approximately 13.3% (Climbfinder, 2025). These climbs are so demanding that they cannot be compared to normal training sessions without them. It could be the effort required of the cyclist for a climb like this that causes them to have an even higher heart rate for the rest of their session. Or, a combination of lesser

climbs that follow each other in these mountainous areas.

To incorporate spatial data into the model, multiple options are available. The data set contains GPS coordinates for every row which could be fed into the model and used as predictors. This could benefit the prediction, but it would be difficult to interpret the results as the variable range is very high. That is why I propose to do a clustering on the GPS coordinates. After that I calculate the distances between the coordinates and each cluster and use them as a predictor in the model. Hierarchical Density-Based Spatial Clustering of Applications with Noise also known as the HDBSCAN algorithm will be used to define clusters in the data (McInnes et al., 2017). This algorithm is chosen because it can handle spatial data which is required in our case. In addition, it is good for handling clusters of varying density. A downside of HDBSCAN is that it is a computationally expensive algorithm, therefore an abstraction had to be made using Uber's Hexagonal Hierarchical Spatial Index (H3), which is a global discrete grid system that indexes geographies on a hexagonal grid (Uber Technologies Inc., 2025). So instead of having, for example, 30 different datapoints on the same road, they belong to the same hexagonal, greatly reducing the amount of data. Popular algorithms such as K-means are not suited for this task because they use Euclidean distance, which is defined as the distance between two points in a Euclidean space, i.e. the shortest path between two points. The Euclidean distance requires a flat surface, hence the reason why it is not suitable here because the dataset consists of geographical coordinates on the Earth's surface, which is spherical rather than planar. That is why the Haversine formula was used:

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

– where φ_1, φ_2 are the latitudes, φ_1, φ_2 the longitudes, and R denotes the radius of the Earth. This formula takes into account the spherical nature of the data, which makes it well suited for this task. The haversine formula was used both for clustering and for calculating the distance between each respective data point and cluster.

3.3 Model

To assess how these variables influence the target variable, multiple machine learning experiments were conducted. For the regression model, XGBoost

was used. XGBoost (Extreme Gradient Boosting) is a tree-based ensemble learning method that can be used for both regression and classification tasks. This model was chosen because of its high performance and accuracy and its ability to scale well to larger datasets. As XGBoost requires careful parameter tuning, hyperparameter optimization was done to obtain the optimal parameters for the model. Bayesian optimization was used because of its high performance and efficiency, which is important since the data set used for training is rather large. For training and evaluation, cross validation was used to prevent overfitting and generalize the model well. If the model was properly trained and could predict the results of the PHR-model, we could conclude that these additional features explain the variance in heart rate by the PHR-model. XGBoost is a high performing model but, due to its complexity, also hard to interpret, a so called 'black box' algorithm. That is why for feature importance SHAP (SHapley Additive exPlanations) was selected (Lundberg & Lee, 2017). This is an explanatory framework for these types of algorithm based on game theory. For the classification experiments, the same model setup was chosen.

4 Experimental setup

4.1 Data

The dataset used for this research consists of three years of race and training data of a cyclist of one of the world tour teams in professional cycling, which is why it is confidential and not publicly available. There are 793 sessions in total, spread throughout the year in different countries. The sessions range from as short as 30 minutes to more than 5 hours. Each session is monitored with a frequency of 1 second with numerous characteristics recorded by a cycling computer, heart strap monitor, and a power meter. Apart from technical attributes that will not benefit this research, the main attributes are:

- Longitude and Latitude
- Left and Right Power pedals (W)
- Cadence (rpm)
- Heart Rate (bpm)

- Speed (km/h)
- Distance (km)
- Altitude (m)
- Riding Time (s)
- Point in time

Since the data contains GPS coordinates and recording dates, the data can be enriched with meteorological data. I have used data from the open-source weather API called Open-Meteo (Open-Meteo, 2025). This API provides historical weather data with a time interval of 1 hour. By extracting every GPS coordinate at the nearest hour from the main data set, I could construct a data set with weather data from every training session and race. This set contains many attributes about the weather, the more important being the following:

- Temperature (°C)
- Dew Point
- Relative Humidity (%)
- Precipitation (mm)
- Evapotranspiration (mm)
- Vapour Pressure Deficit (kPa)
- Surface pressure (Pa)

As the cycling data set contains data with a frequency of 1 second, it was not possible to accurately obtain the weather at every moment. That is why the assumption had to be made that the weather variables do not change dramatically in 30 minutes. This means, for example, that data ranging from 09:30:01 to 10:30:00 have the same weather data as the data at 10:00:00. In this manner, the data set contains weather data at all times and can be used for modeling.

Group	Description	Number of Features
Geo	Distance-based features derived from the clustering experiment.	44
Meteo	Meteorological variables including temperature, wind speed and relative humidity.	14
Parcours	Terrain and riding characteristics including slope, altitude, cadence, speed and duration features.	15

Table 1: Definition of feature groups used for modeling.

Parameter	Value
Minimum cluster size (`min_cluster_size`)	1200
Minimum samples (`min_samples`)	1200
Distance metric	Haversine

Table 2: HDBSCAN hyperparameters used for clustering experiment.

4.2 Feature groups

In addition to the features already mentioned in Chapter 4.1, multiple features were constructed to improve modeling. For a complete list of all features, see Chapter 7. To see how each different type of feature contributed to the model, the features were categorized into groups based on their type.

Each feature group in Table 1 is evaluated separately as well as in any possible combination of feature groups.

4.3 Model setup

The H3 grid system was applied at resolution 10, corresponding to hexagonal cells with an average area of $0.01505, \mathrm{km}^2$ and an average edge length of $0.07586, \mathrm{km}$. This resolution was chosen to maintain a high level of spatial discrimination while remaining computationally stable.

The HDBSCAN parameters in Table 2 were such that the output of the clusters was limited. This is because too many clusters would have a negative impact on the model and high computational cost. For hyperparameter

Parameter	Value
Objective	`reg:squarederror`
Number of trees (`n_estimators`)	294
Maximum tree depth (`max_depth`)	10
Learning rate (`learning_rate`)	0.1794
Subsample ratio (`subsample`)	0.6231
Column subsample ratio (`colsample_bytree`)	0.9020
Minimum child weight (`min_child_weight`)	2
Tree construction method	`hist`

Table 3: XGBoost hyperparameters of the regression experiment.

Parameter	Value
Objective	`binary:logistic`
Number of trees (`n_estimators`)	299
Maximum tree depth (`max_depth`)	10
Learning rate (`learning_rate`)	0.1983
Subsample ratio (`subsample`)	0.7719
Column subsample ratio (`colsample_bytree`)	0.7142
Minimum child weight (`min_child_weight`)	6
Tree construction method	`hist`

Table 4: XGBoost hyperparameters used in the classification experiment.

tuning of both classification and regression experiments, the Optuna framework was used to perform Bayesian optimization (Akiba et al., 2019). For all combinations of feature groups, the same set of hyperparameters was used for each respective model. Regression performance was evaluated using R^2, root mean squared error (RMSE), and mean absolute error (MAE) under a 10-fold cross-validation, with all metrics reported as averages in the validation folds. Classification performance was evaluated using the area under the receiver operating characteristic curve (ROC–AUC) under a 10-fold cross-validation. Accuracy, precision, recall, and F1-score were additionally reported as averages across validation folds to provide complementary information on classification performance.

Following model training, feature contributions were analyzed using SHAP.

SHAP requires a trained model and a representative background dataset. As training ride sessions vary considerably between locations and seasons, a stratified sample of the entire dataset was used to ensure that the data from all sessions were proportionally represented. Since SHAP uses the model trained on the complete data set, this subset of data is sufficient to define the importance of the features.

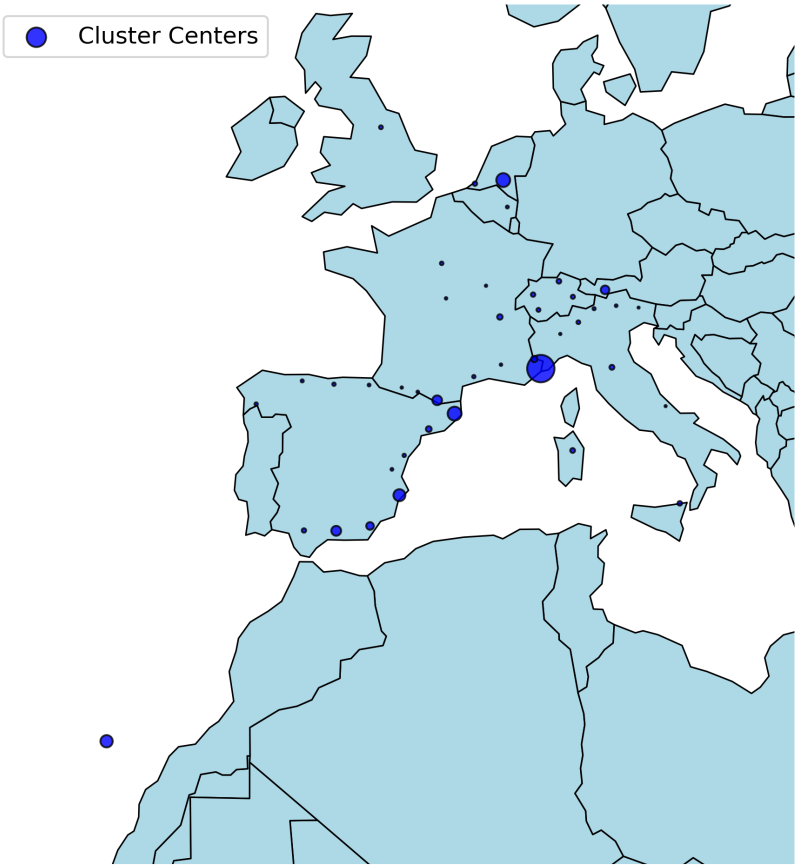


Figure 4: Results of the HDBSCAN clustering experiment.

5 Results

5.1 Clustering and regression

So before the regression experiment and classification could be conducted the distance features need to be created. Figure 4 shows that the clusters are located mainly in Europe and one in Tenerife, there were also two clusters in Brazil and United Arab Emirates which were excluded for display purposes. Furthermore, it is notable that the larger clusters are in mountainous area's near Monaco and the Spanish coast and Pyrenees. With each cluster defined, distance features could be calculated for every data-point.

Table 5 shows that the combination of all features had the best in R^2

Geo	**Meteo**	**Parcours**	R^2	**RMSE**	**MAE**
1	0	0	0.4205	5.1120	3.6475
0	1	0	0.3683	5.3376	3.7756
0	0	1	0.6740	3.8341	2.7938
1	1	0	0.6817	3.7887	2.5819
1	0	1	0.7425	3.4080	2.4611
0	1	1	0.7767	3.1731	**2.2642**
1	**1**	**1**	**0.7768**	**3.1727**	2.2686

Table 5: Results of the cross-validated regression experiment for all evaluated feature combinations.

and RMSE, but a slightly higher MAE than the combination of Meteo and Parcours features. An R^2 of 0.7768, that is, 77.67% of the variance of the residual heart rate is explained by the model, is considered relatively good. The model achieves a mean absolute error (MAE) of 2.2686 and a root mean squared error (RMSE) of 3.1727. The MAE indicates that on average the predicted residual deviate slightly from the actual residual, showing some high predictive accuracy. The fact that the RMSE is higher relative to the MAE suggests that the model occasionally has larger prediction errors. It is noticeable that the combination of all features just slightly outperforms the second best combination. When independently evaluated, the Parcours feature group provides the strongest baseline performance. In this setting, the Geo features outperform the Meteo features. However, when combined with the Parcours features, the Geo features do not yield additional performance gains compared to the Meteo feature group combined with the Parcours feature group. A possible explanation for this can be that the Geo and Parcours features explain the same variance of the PHR model compared to the Meteo features.

Looking solely at the Geo features, the experiment has a R^2 of 0.419, which means that the current features explain 41,9% of the variance in the residuals. This is not a high score, but it was expected because the model only used a subset of the features. However, it does mean that the distance measures have some correlation with the residuals.

In Table 6 the results of the feature importance of SHAP are shown. Cluster 1 is the best performing feature that has an average SHAP value of 0.3693 impact on the residual, which means that on average the model

Distance feature	Mean absolute SHAP value	Latitude	Longitude	Area type
1	0.3693	28.27	-16.60	Mountainous
2	0.3205	48.21	1.94	Flat
3	0.2885	38.54	-0.40	Mountainous
4	0.2880	43.98	5.22	Mountainous
5	0.2880	37.05	-3.89	Mountainous
6	0.2625	37.06	-5.68	Mountainous
7	0.2565	53.89	-1.42	Hilly
8	0.2429	43.30	-5.77	Mountainous
9	0.2414	44.22	7.06	Mountainous
10	0.2364	37.25	-2.02	Mountainous

Table 6: Top 10 best performing distance features based on mean absolute SHAP value.

adds or subtracts 0.3693 to the target variable. This only tells something about the importance of the feature, but not necessarily why this feature is important, i.e., are training rides near cluster 1 causing higher or lower residuals? In Figure 5 you can see how the distance to cluster 1 impacts the SHAP value and thus the model. It is especially interesting that when the distance is close to 0, the SHAP values are generally higher than the average in absolute terms. This means that the residuals are higher when the training rides are near cluster 1.

In Figure 6 you can see the important features and that cluster 1 is in Tenerife. Something in Tenerife, like the climate or the numerous different climbs on the island, causes the PHR-model to have larger prediction errors. The next best performing cluster 2 is in Toury, France. This is notable because this cluster is not as large as other clusters, but still has a relatively big influence on the model. The dependency plot of cluster 2 in Figure 7 shows that the SHAP values are almost only lower when near this cluster. Since this is the only cluster with a flat area type, this could mean that there is a correlation between the target variable and the area types.

It can be said that the cluster distances have some impact on the model looking at the variance that it explains. However, from this feature group only it is hard to draw any factual conclusions on what the causing factors behind these clusters impact the model. The initial hypothesis that certain

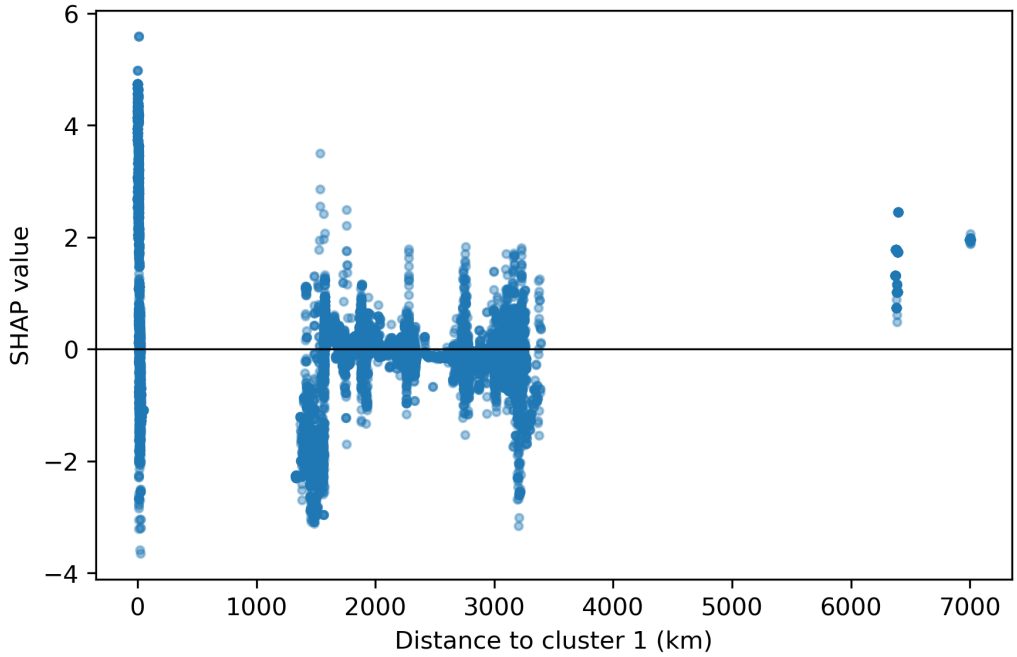


Figure 5: SHAP dependency plot for distance cluster 1.

climbs have a great impact on the model cannot be solely proven based on these results. The clustering results show clusters on all sorts of area's and when looking at the more important clusters, the clusters generally are in mountainous area's. Still, it is not enough to draw a conclusion, since there are many variables tied to an area which can have an effect on the model. In addition to the climbs, the climate in these areas ranges from colder in northern France to warm and windy in Tenerife, which can be important, as previously discussed. Furthermore, the data set contains rides from different periods in time of one single rider. It could be that the rider was in a different kind of shape or that a new sensor was used to monitor the rides.

Since the combination of all features is the best performing set, it is good to dive deeper into how it performs. Figure 8 shows a strong alignment of the predicted residuals and the observed residual, which can be seen from the high density of bins on the diagonal. This indicates that the model is properly trained on the majority of the data. The highest density occurs around residuals that are zero or close to zero, indicated by the yellow hexagonal. The data set contains mostly residuals close to zero which the model seems

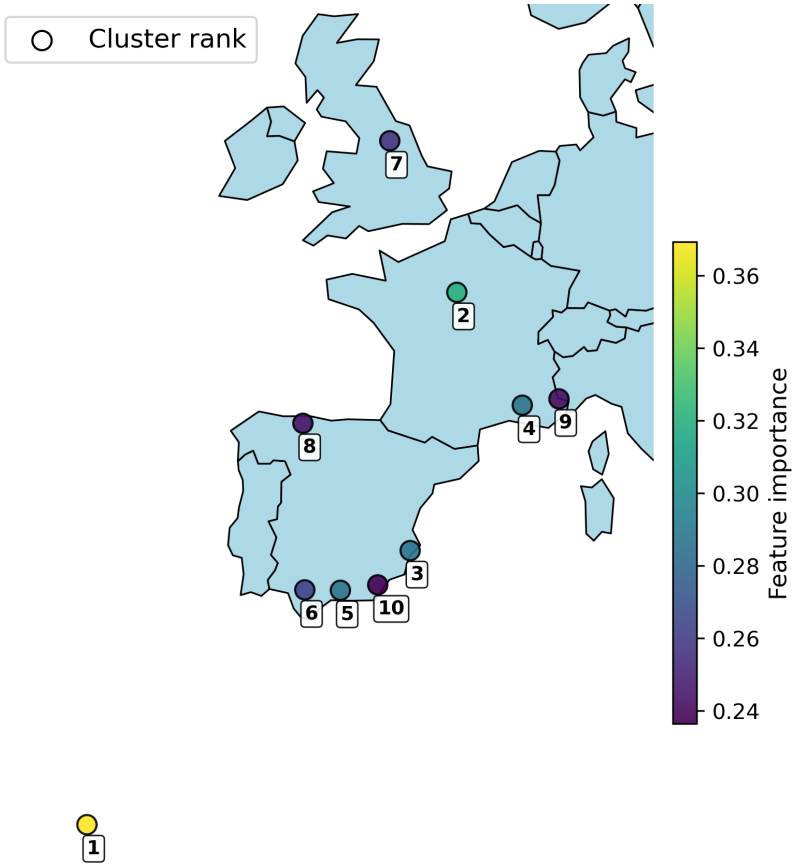


Figure 6: Graphical presentation of the 10 best performing clusters.

to predict well. The model handles larger positive and negative residuals differently. Although extreme residuals are inherently more difficult to predict, the model seems to be able to predict negative residuals. The positive ones, on the other hand, the model does not seem to predict as good and tends to predict lower residuals than the actual residuals.

The bar graph in Figure 9 shows that the riding time is the most important feature in predicting the heart rate residuals, closely followed by the distance traveled in a ride. This could possibly be explained by the cardiovasculair drift that after prolonged exercise the heart race increases at the

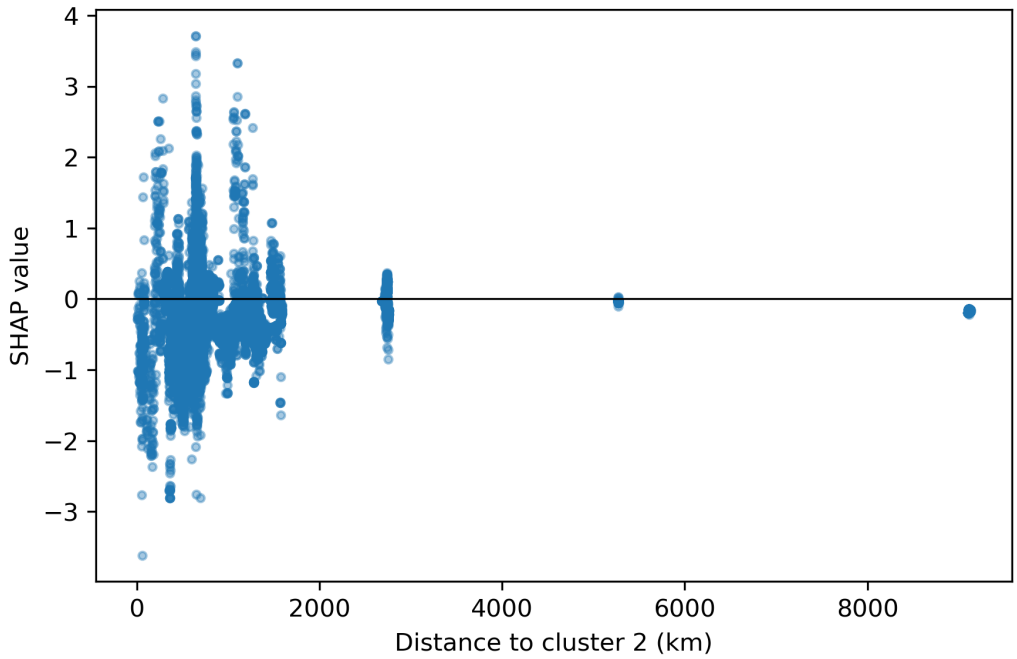


Figure 7: SHAP dependency plot for distance cluster 2.

same exertion (Coyle & Gonzalez-Alonso, 2001). On the beeswarm graph this is confirmed when you see that when the feature value of the riding time is high, the SHAP values are also increasingly higher. When the riding time is low, the model regularly predicts a lower residual than when it is higher. The PHR-model fits a sigmoid curve on the data of a complete single ride. Taking into account the effects of cardiovasculair drift, the model would underestimate the heart rate at the beginning of each ride, hence the reason for the large blue cluster in the riding time feature.

If riding time is the best performing feature, it makes sense that distance and cumulative gained altitude meters are the second and third best performing features, since these features are highly correlated to each other. When riding time increases, it is to be expected that riding distance increases, and in a mountainous area the gained altitude meters also increase. This is confirmed by the Pearson correlation test, which shows that riding distance and cumulative gain were strongly positively correlated (Pearson's $r = 0.78$, $p < 0.001$). Further alignment can be observed in Figure 10. As riding time increases, the residuals tend to increase. At the same time, the distance

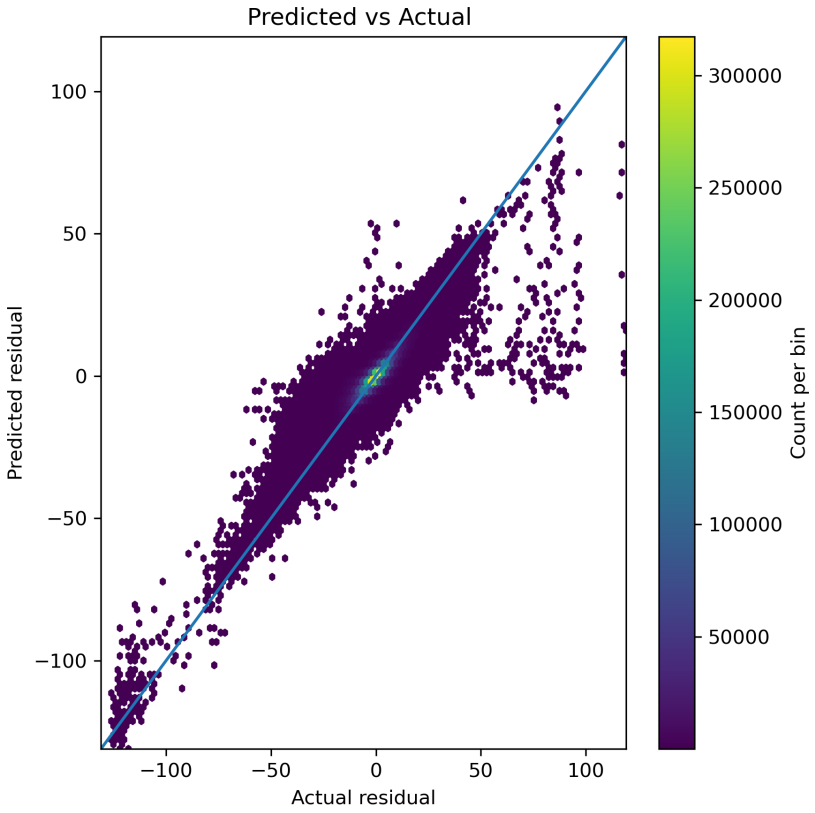


Figure 8: Scatter plot of the predicted residual versus the actual residual.

traveled also increases, as indicated by the red-colored data points.

The results of these features are roughly the same as with riding time, where a lower value has a near zero or negative value, and the larger feature values have a higher SHAP value.

Higher temperature has a negative effect on prediction, which is surprising because research showed that increasing temperature can increase heart rate at the same intensity. In Figure 11 it is seen that there is a downward trend in SHAP values when the temperature increases. A possible explanation for this can be that the temperature usually stays the same during training rides, so the model does not really capture the influence of the change in temperature. It can be that the average heart rate of one ride is higher compared to another ride in a more cold climate, but that does not mean that the residuals are higher since each prediction error comes from a model

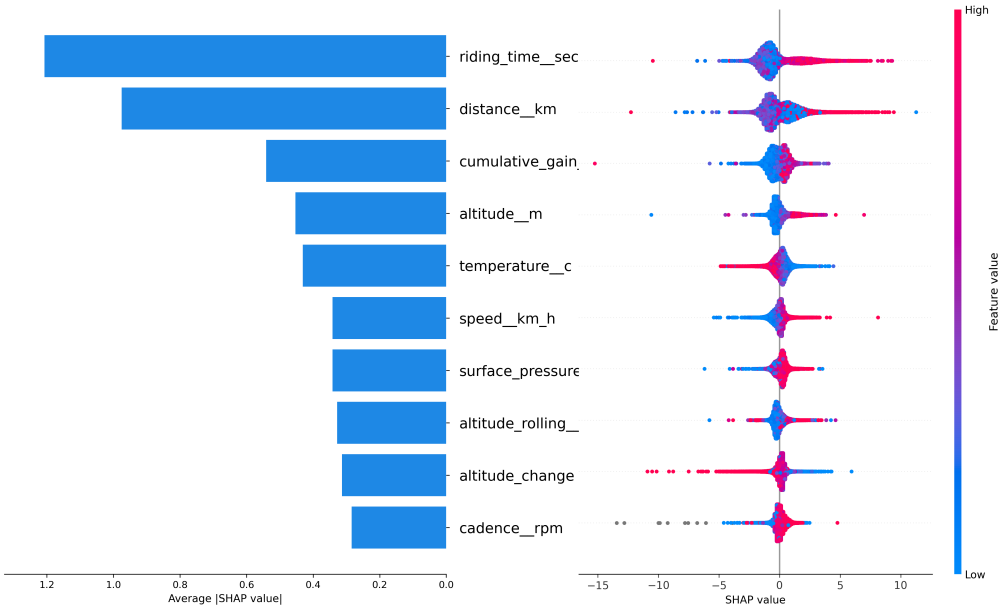


Figure 9: Feature importance plot of the regression experiment. The top 10 best performing features are shown. On the left the bar plot shows impact of the features in terms of magnitude and on the right how these features impact the prediction with their respective value.

based on a single ride. If the PHR-model was constructed based on multiple rides, you would probably see that the residuals would be higher in warmer climates as long as the training rides were in different climates.

Moreover, Figure 9 identifies surface pressure as another Meteo feature associated with higher residual values. Other Meteo variables do not appear to have a strong predictive influence, and none of the Geo features are present among the top ten most important predictors. A complete overview of the feature importance results for the regression experiment is provided in Chapter 7.

5.2 Classification

Looking at the results in Table 7, the classification model performed well with an ROC AUC score of 0.9919, which means that the model is highly effective in ranking positive instances above negative ones in all possible classification

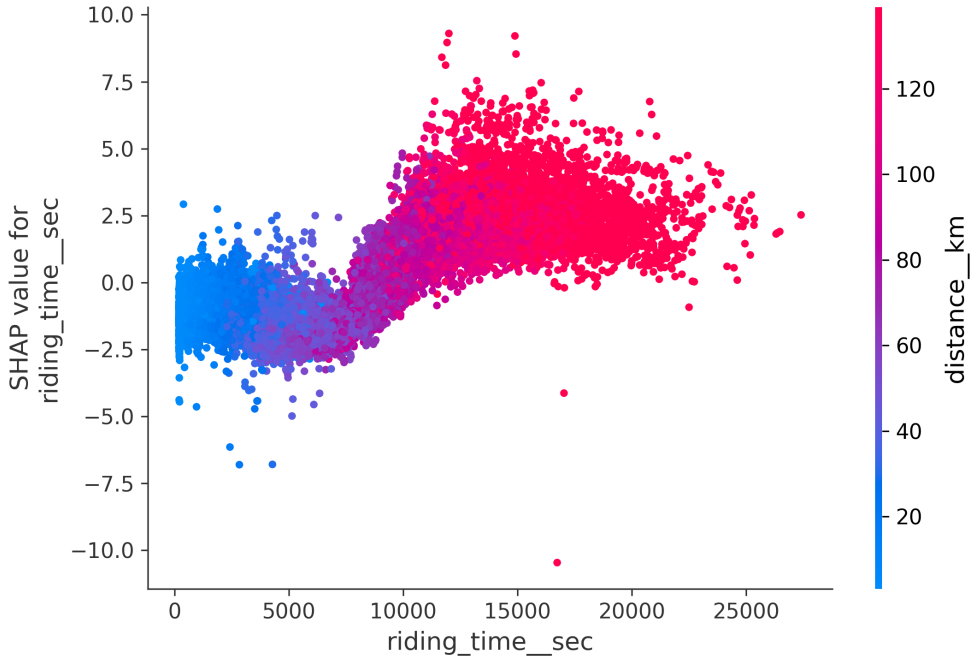


Figure 10: SHAP dependency plot for the Riding time feature of the regression experiment. The coloring shows how riding time interacts with distance.

thresholds. The PR AUC is more focused on positive cases and penalizes false positives more, which is important in this case because the data set used for this experiment was imbalanced. With a score of 0.9101, the model also handles efficiently when it is more emphasized on positive cases, which can also be concluded when looking at the precision of 0.8918. The model is 89,18% of the time correct when it predicts a positive target. However, with a recall of 0.7522, meaning that the model misses 24,88% of the positive cases, there is still room for improvement. This combined gives an F1-score of 0.8161, which is considered good but not great.

This is further highlighted by the ROC curve shown in Figure 12, which is in line with the results mentioned above. The curve rises steeply from the origin and approaches a true positive rate close to 1.0 at very low false positive rates, indicating that the model is good at distinguishing between positive and negative cases. Across a wide range of false positive rate thresholds, the classifier achieves a high true positive rate, which is reflected in the near-

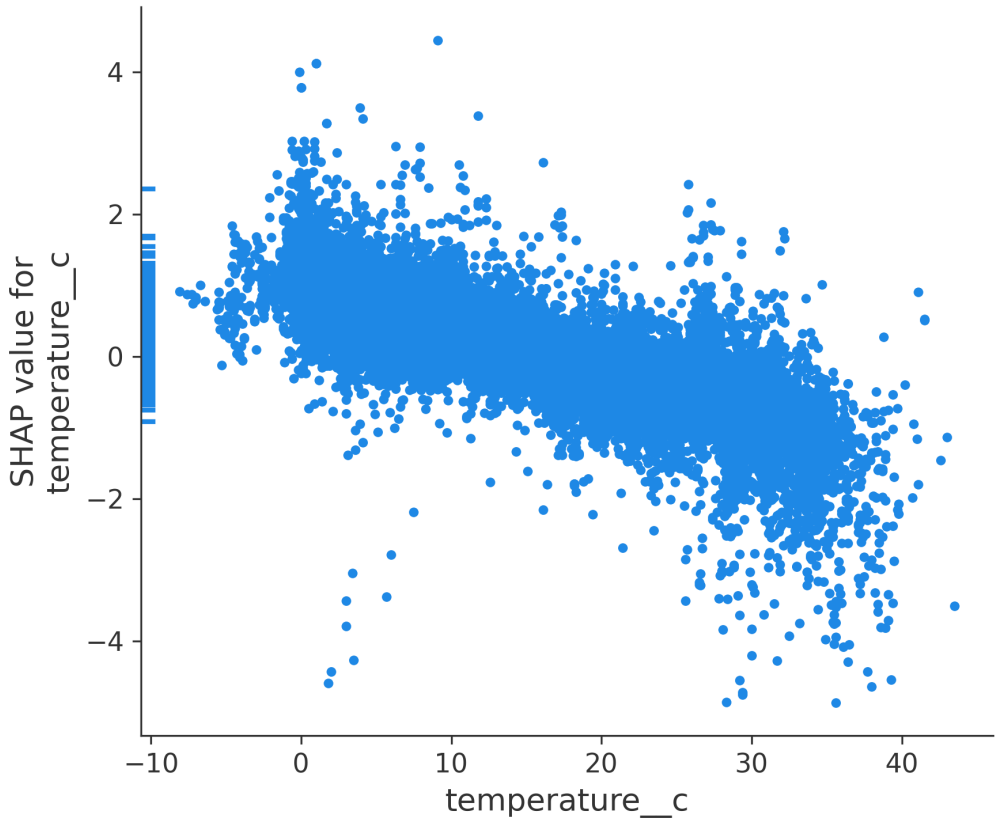


Figure 11: SHAP dependency plot for the Temperature feature of the regression experiment.

Metric	**Value CV mean**
ROC AUC	0.9919
PR AUC	0.9101
Precision	0.8918
Recall	0.7522
F1-score	0.8161

Table 7: Classification performance statistics.

perfect ROC AUC. This confirms the strong discriminative performance of the classification model.

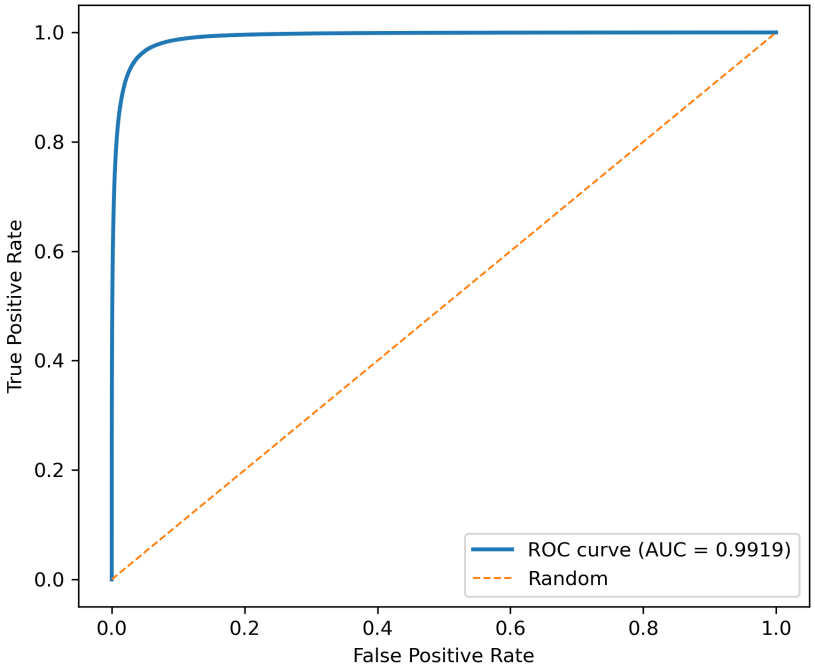


Figure 12: AUC ROC curve of the classification experiment.

Figure 13 shows that for the classification model, the distance traveled is the best performing feature, closely followed by the cumulative gained altitude meters. When cumulative gain and distance ridden are high, the model tends to give a higher SHAP value, e.g. predicts towards a positive class. Similarly to the regression experiment, all these features have something to do with prolonged exercise and thus cardiovasculair drift, resulting in a higher heart heart rate and residual.

Furthermore, it is worth noting that the high altitude change, i.e. the changed altitude in 30 seconds, has a large negative impact on the prediction when the feature value is high. So if you are ascending, model tends to predict a negative class. The altitude rolling shows that when the average altitude of the last 30 seconds is high, the model predicts a positive class. Being longer at high altitudes is correlated to more positive instances.

As mentioned earlier, the regression experiment looked at every residual of the PHR-model, while the classification experiment had a threshold of minimum 10 to be considered a target. It is interesting to see if there are any differences between these two experiments. Looking at Figures 9 and 13,

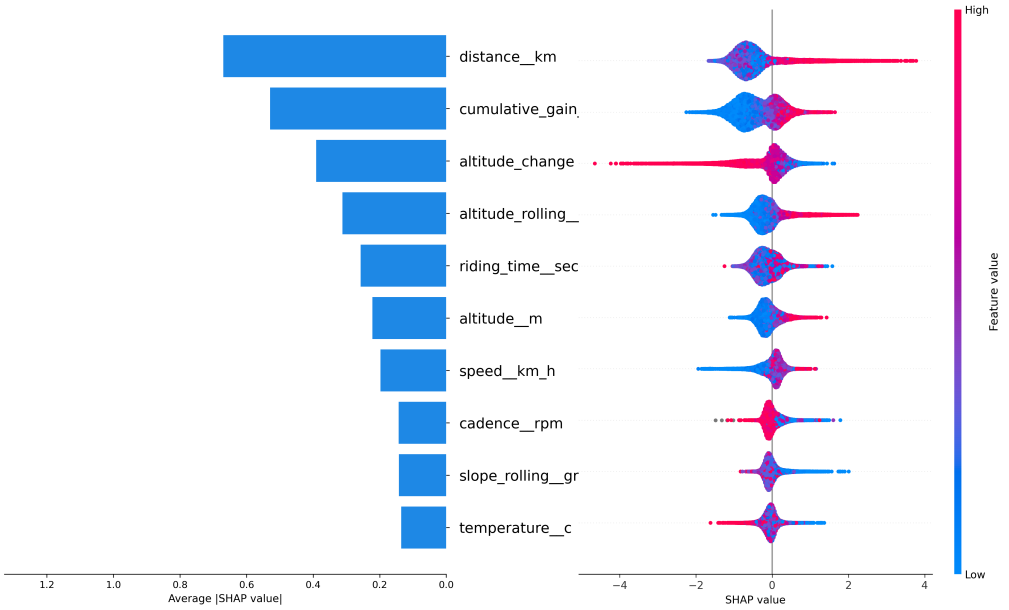


Figure 13: Feature importance plot of the classification experiment. The top 10 best performing features are shown. On the left the bar plot shows impact of the features in terms of magnitude and on the right how these features impact the prediction with their respective value.

the top 10 performing features are the same for both experiments, except for order and magnitude. The first thing noticeable is the different top performing feature. The cumulative gain is in the regression experiment third-best and a much lower impact than the number one ride time. In classification, cumulative gain is the second best performing feature, although by a smaller margin than in the regression experiment. A possible explanation could be that training sessions that cover many altitude meters are usually harder for the rider, causing a much more fatigued body and a higher heart rate for the same exertion. Riding time also seems to lose some of his predictive power in the classifcation experiment as it has a much lower impact, being the 5th best performing feature. As the magnitudes of the different SHAP values cannot be compared between the two experiments since the target values are different, they can be compared to their own experiment. Then it is noticeable that in the regression experiment the best performing feature riding time has an absolute impact twice as big as the third feature cumulative gain. How-

ever, in the classification experiment, the features are much closer to each other and riding time is not even the best performing feature. A detailed and complete overview of the importance of the classification experiment features can be found in the Chapter 7 as well.

6 Discussion

In this report I conducted research on the Power Heart-rate model for continuous fitness monitoring in cycling provided by (de Leeuw et al., 2023). This model creates a direct relationship between power output and heart rate, and by fitting a sigmoid curve on these data, one could say something about the fitness level of a cyclist. As the model shows inefficiencies in terms of the differences between the curve and the predicted heart rate, other factors could possibly have an effect on the heart rate that goes beyond the power output. The goal of this research was to discover which kind of variables affect the heart rate in order to improve model performance and thus assessing a rider's fitness. I did this by conducting three machine learning experiments to discover these variables.

The first experiment researched the effects of GPS data on the target variable. Since raw GPS coordinates are not ideal for machine learning tasks, I did a HDBSCAN clustering to identify clusters across the data set. With the clusters I created distance features for every row of the dataset with distance measured of each respective cluster. The second experiment was a regression task that included feature groups, showing that the combination of all features has the best results, as seen in Table 5. These results clearly indicated that the features have a prediction value for the residuals. Figure 9 showed the importance of each feature and their contribution to the model. The third experiment was a classification experiment. In this experiment, the model was good in distinguishing positive and negative cases as seen in Table 7. The feature importance in Figure 13 showed that the same features were important but that the order and magnitude were different compared to the regression. In conclusion of these experiments, especially variables that are correlated to the length of a session, such as distance ridden, cumulative gain altitude meters or just the time of a session, have a great impact on the PHR-model. In addition, weather variables such as temperature and surface pressure are important, but distance features seem to have a smaller effect if all features are taken into account. Looking at the feature importance's from the different experiments, it is important to note that this does not necessarily define a causal relationship between the residuals and the features. The feature importance are designed to show correlation between the predictor variables and target variable but without causation. This is because variables can have their own causing factors, which might lead to the high performance of this feature and not necessarily the feature itself. It is also the

reason why the cluster distances and their performance can say something about the effects of the location but no causation. The training rides are in many different locations during different periods of the year. So, it could be that some location had bad weather conditions causing high residuals, instead of difficult bike courses. For further research, these variables should be incorporated into the model in order to improve performance.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W., et al. (2017). Monitoring athlete training loads: Consensus statement. *International journal of sports physiology and performance*, *12*(s2), S2–161.

Climbfinder. (2025). Climbfinder.

Coyle, E. F., & Gonzalez-Alonso, J. (2001). Cardiovascular drift during prolonged exercise: New perspectives. *Exercise and sport sciences reviews*, *29*(2), 88–92.

de Leeuw, A.-W., Heijboer, M., Verdonck, T., Knobbe, A., & Latré, S. (2023). Exploiting sensor data in professional road cycling: Personalized data-driven approach for frequent fitness monitoring. *Data Mining and Knowledge Discovery*, *37*(3), 1125–1153.

Fry, R. W., Morton, A. R., & Keast, D. (1991). Overtraining in athletes: An update. *Sports medicine*, *12*(1), 32–65.

Galloway, S., & Maughan, R. J. (1997). Effects of ambient temperature on the capacity to perform prolonged cycle exercise in man. *Medicine and science in sports and exercise*, *29*(9), 1240–1249.

Hirschman, I. I., & Widder, D. V. (2012). *The convolution transform*. Courier Corporation.

Joyner, M. J., & Coyle, E. F. (2008). Endurance exercise performance: The physiology of champions. *The Journal of physiology*, *586*(1), 35–44.

Kellmann, M. (2010). Preventing overtraining in athletes in high-intensity sports and stress/recovery monitoring. *Scandinavian journal of medicine & science in sports*, *20*, 95–102.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

McGrath, E., Mahony, N., Fleming, N., & Donne, B. (2019). Is the ftp test a reliable, reproducible and functional assessment tool in highly-trained athletes? *International journal of exercise science*, *12*(4), 1334.

McInnes, L., Healy, J., Astels, S., et al. (2017). Hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

Nazaret, A., Tonekaboni, S., Darnell, G., Ren, S. Y., Sapiro, G., & Miller, A. C. (2023). Modeling personalized heart rate response to exercise and environmental factors with wearables data. *NPJ Digital Medicine*, *6*(1), 207.

Open-Meteo. (2025). Open-meteo weather forecast api.

Taelman, J., Vandeput, S., Spaepen, A., & Van Huffel, S. (2009). Influence of mental stress on heart rate and heart rate variability. *4th European Conference of the International Federation for Medical and Biological Engineering: ECIFMBE 2008 23–27 November 2008 Antwerp, Belgium*, 1366–1369.

Uber Technologies Inc. (2025). H3: Uber's hexagonal hierarchical spatial index.

7 Appendix

Feature	**Description**	**SHAP**	**SHAP (norm.)**
riding_time__sec	Ride duration (s)	1.2070	1.0000
distance__km	Ride distance (km)	0.9749	0.8072
cumulative_gain__km	Cumulative elevation gain (km)	0.5408	0.4465
altitude__m	Altitude (m)	0.4528	0.3734
temperature__c	Air temperature (°C)	0.4313	0.3555

Feature	Description	SHAP	SHAP (norm.)
speed__km_h	Speed (km/h)	0.3417	0.2811
surface_pressure__pa	Surface pressure (Pa)	0.3415	0.2809
altitude_rolling__m	Rolling altitude (m)	0.3276	0.2694
altitude_change	Altitude change	0.3131	0.2573
cadence__rpm	Cadence (rpm)	0.2843	0.2334
temperature_2m	Air temperature at 2 m (°C)	0.2512	0.2059
apparent_temperature	Apparent temperature (°C)	0.2470	0.2024
slope_rolling__grad	Rolling slope (°)	0.2012	0.1643
dist_km_c23	Distance from cluster 23	0.1978	0.1615
relativehumidity_2m	Relative humidity (%)	0.1952	0.1594
dist_km_c20	Distance from cluster 20	0.1869	0.1524
dist_km_c15	Distance from cluster 15	0.1752	0.1427
wind_speed_10m	Wind speed at 10 m (m/s)	0.1716	0.1398
wind_direction_10m	Wind direction at 10 m (°)	0.1653	0.1345
dew_point_2m	Dew point (°C)	0.1584	0.1288
dist_km_c4	Distance from cluster 4	0.1583	0.1287
cloudcover_mid	Mid cloud cover (%)	0.1559	0.1267
dist_km_c14	Distance from cluster 14	0.1355	0.1097
dist_km_c9	Distance from cluster 9	0.1318	0.1067
dist_km_c13	Distance from cluster 13	0.1291	0.1044
dist_km_c24	Distance from cluster 24	0.1288	0.1041
dist_km_c19	Distance from cluster 19	0.1280	0.1035
dist_km_c42	Distance from cluster 42	0.1259	0.1017
slope_change	Slope change	0.1240	0.1002
dist_km_c16	Distance from cluster 16	0.1197	0.0966
cloudcover	Total cloud cover (%)	0.1195	0.0964
cloudcover_low	Low cloud cover (%)	0.1156	0.0932
altitude_diff	Altitude difference (m)	0.1140	0.0919
slope	Slope (°)	0.1123	0.0905
dist_km_c7	Distance from cluster 7	0.1114	0.0897
dist_km_c41	Distance from cluster 41	0.1059	0.0852
dist_km_c21	Distance from cluster 21	0.1053	0.0846
cloudcover_high	High cloud cover (%)	0.1036	0.0832
dist_km_c2	Distance from cluster 2	0.0985	0.0790
dist_km_c8	Distance from cluster 8	0.0979	0.0785
dist_km_c43	Distance from cluster 43	0.0965	0.0773

Feature	Description	SHAP	SHAP (norm.)
dist_km_c6	Distance from cluster 6	0.0959	0.0768
dist_km_c37	Distance from cluster 37	0.0949	0.0760
dist_km_c36	Distance from cluster 36	0.0935	0.0748
dist_km_c5	Distance from cluster 5	0.0933	0.0747
dist_km_c33	Distance from cluster 33	0.0928	0.0743
dist_km_c25	Distance from cluster 25	0.0910	0.0728
dist_km_c30	Distance from cluster 30	0.0900	0.0719
dist_km_c12	Distance from cluster 12	0.0887	0.0708
dist_km_c27	Distance from cluster 27	0.0885	0.0707
dist_km_c1	Distance from cluster 1	0.0883	0.0706
dist_km_c28	Distance from cluster 28	0.0864	0.0689
dist_km_c22	Distance from cluster 22	0.0850	0.0678
dist_km_c29	Distance from cluster 29	0.0849	0.0677
dist_km_c11	Distance from cluster 11	0.0829	0.0661
dist_km_c26	Distance from cluster 26	0.0810	0.0645
dist_km_c31	Distance from cluster 31	0.0802	0.0638
dist_km_c18	Distance from cluster 18	0.0775	0.0615
dist_km_c10	Distance from cluster 10	0.0747	0.0593
dist_km_c40	Distance from cluster 40	0.0687	0.0543
dist_km_c0	Distance from cluster 0	0.0647	0.0509
dist_km_c39	Distance from cluster 39	0.0643	0.0506
dist_km_c38	Distance from cluster 38	0.0642	0.0505
dist_km_c3	Distance from cluster 3	0.0635	0.0499
dist_km_c17	Distance from cluster 17	0.0620	0.0487
dist_km_c35	Distance from cluster 35	0.0584	0.0457
dist_km_c32	Distance from cluster 32	0.0509	0.0395
precipitation	Precipitation (mm)	0.0433	0.0331
dist_km_c34	Distance from cluster 34	0.0419	0.0320
weathercode	Weather code	0.0295	0.0217
sunshine_duration	Sunshine duration	0.0232	0.0164
rain	Rain indicator	0.0113	0.0066
slope_percent	Slope (%)	0.0093	0.0049
elevation_gain	Elevation gain (m)	0.0064	0.0025
distance_diff	Distance change (km)	0.0034	0.0000

Table 8: Feature results of the regression experiment

Feature	Description	SHAP	SHAP (norm.)
distance__km	Ride distance (km)	0.6696	1.0000
cumulative_gain__km	Cumulative elevation gain (km)	0.5291	0.7891
altitude_change	Altitude change	0.3909	0.5816
altitude_rolling__m	Rolling altitude (m)	0.3122	0.4633
riding_time__sec	Ride duration (s)	0.2575	0.3812
altitude__m	Altitude (m)	0.2219	0.3277
speed__km_h	Speed (km/h)	0.1978	0.2916
cadence__rpm	Cadence (rpm)	0.1431	0.2095
slope_rolling__grad	Rolling slope (°)	0.1421	0.2079
temperature__c	Air temperature (°C)	0.1351	0.1975
surface_pressure__pa	Surface pressure (Pa)	0.1268	0.1849
relativehumidity_2m	Relative humidity (%)	0.1149	0.1671
temperature_2m	Air temperature at 2 m (°C)	0.1063	0.1543
apparent_temperature	Apparent temperature (°C)	0.1041	0.1509
wind_speed_10m	Wind speed at 10 m (m/s)	0.1025	0.1485
cloudcover_low	Low cloud cover (%)	0.0945	0.1366
dist_km_c1	Distance from cluster 1	0.0896	0.1292
dew_point_2m	Dew point (°C)	0.0866	0.1247
dist_km_c18	Distance from cluster 18	0.0849	0.1220
cloudcover	Total cloud cover (%)	0.0825	0.1185
wind_direction_10m	Wind direction at 10 m (°)	0.0819	0.1176
cloudcover_mid	Mid cloud cover (%)	0.0809	0.1161
dist_km_c23	Distance from cluster 23	0.0746	0.1066
dist_km_c5	Distance from cluster 5	0.0732	0.1045
dist_km_c4	Distance from cluster 4	0.0728	0.1038
dist_km_c31	Distance from cluster 31	0.0726	0.1037
cloudcover_high	High cloud cover (%)	0.0703	0.1002
dist_km_c22	Distance from cluster 22	0.0693	0.0987
dist_km_c25	Distance from cluster 25	0.0687	0.0978
dist_km_c14	Distance from cluster 14	0.0680	0.0967
dist_km_c43	Distance from cluster 43	0.0675	0.0959
dist_km_c20	Distance from cluster 20	0.0666	0.0946
dist_km_c15	Distance from cluster 15	0.0655	0.0929
dist_km_c24	Distance from cluster 24	0.0654	0.0928
dist_km_c2	Distance from cluster 2	0.0650	0.0922
dist_km_c8	Distance from cluster 8	0.0622	0.0881

Feature	Description	SHAP	SHAP (norm.)
dist_km_c0	Distance from cluster 0	0.0582	0.0820
dist_km_c6	Distance from cluster 6	0.0565	0.0794
dist_km_c30	Distance from cluster 30	0.0516	0.0721
dist_km_c36	Distance from cluster 36	0.0505	0.0704
dist_km_c21	Distance from cluster 21	0.0501	0.0698
dist_km_c10	Distance from cluster 10	0.0499	0.0695
dist_km_c42	Distance from cluster 42	0.0495	0.0690
dist_km_c39	Distance from cluster 39	0.0495	0.0689
dist_km_c19	Distance from cluster 19	0.0481	0.0668
dist_km_c26	Distance from cluster 26	0.0478	0.0663
dist_km_c29	Distance from cluster 29	0.0476	0.0661
dist_km_c28	Distance from cluster 28	0.0464	0.0642
dist_km_c9	Distance from cluster 9	0.0460	0.0637
dist_km_c37	Distance from cluster 37	0.0446	0.0616
dist_km_c7	Distance from cluster 7	0.0444	0.0613
dist_km_c16	Distance from cluster 16	0.0436	0.0601
dist_km_c11	Distance from cluster 11	0.0435	0.0599
dist_km_c3	Distance from cluster 3	0.0411	0.0563
dist_km_c32	Distance from cluster 32	0.0407	0.0557
dist_km_c40	Distance from cluster 40	0.0405	0.0554
dist_km_c17	Distance from cluster 17	0.0396	0.0540
dist_km_c12	Distance from cluster 12	0.0389	0.0530
slope_change	Slope change	0.0388	0.0529
dist_km_c34	Distance from cluster 34	0.0387	0.0526
dist_km_c41	Distance from cluster 41	0.0380	0.0517
dist_km_c13	Distance from cluster 13	0.0348	0.0468
dist_km_c27	Distance from cluster 27	0.0330	0.0442
elevation_gain	Elevation gain (m)	0.0303	0.0401
dist_km_c33	Distance from cluster 33	0.0283	0.0371
altitude_diff	Altitude difference (m)	0.0277	0.0362
dist_km_c35	Distance from cluster 35	0.0274	0.0358
dist_km_c38	Distance from cluster 38	0.0220	0.0276
precipitation	Precipitation (mm)	0.0168	0.0198
sunshine_duration	Sunshine duration	0.0162	0.0190
weathercode	Weather code	0.0149	0.0170
slope	Slope (°)	0.0148	0.0168

Feature	**Description**	**SHAP**	**SHAP (norm.)**
rain	Rain indicator	0.0079	0.0065
slope_percent	Slope (%)	0.0071	0.0053
distance_diff	Distance change (km)	0.0036	0.0000

Table 9: Feature results of the classification experiment