Data Mining Project

University of Pisa - Computer Science
Nicola Emmolo
Simone Marzeddu
Jacopo Raffi

Academic Year 2024/2025

Contents

1 Introduction

This report details a comprehensive data mining project aimed at analyzing the dynamics of cycling competitions and the attributes of cyclists. The project explores various stages of data processing and analysis, leveraging a blend of modern machine learning techniques and statistical methodologies. The analysis is based on two datasets: one focused on cyclist characteristics and the other on race-specific information. These datasets were examined and combined to enable a holistic exploration of factors affecting performance outcomes in competitive cycling.

The project begins with a data understanding phase, where syntactic and semantic evaluations were conducted to ensure dataset quality. Attributes were analyzed for their distributions, correlations, and potential biases. Subsequently, the preprocessing phase addressed challenges like missing data and inconsistencies through strategies such as feature imputation, outlier detection using Isolation Forest, and dimensionality reduction with PCA and UMAP. Feature engineering played a pivotal role, introducing attributes to better capture the nuances of the data.

Clustering methods, including DBSCAN, hierarchical clustering, and K-Means, were then applied to group data into meaningful clusters. These clusters provided insights into the segmentation of cyclists and races based on performance metrics and race characteristics. Classification models were employed to predict top-20 placements, utilizing engineered features and balancing techniques to address dataset imbalances.

In the final stage, explainability techniques like SHAP, LIME, and LORE were used to interpret the models' decisions, providing transparency and revealing limitations that could guide future improvements. Through these steps, the project aims at offering a detailed framework for understanding and predicting outcomes in the domain of competitive cycling.

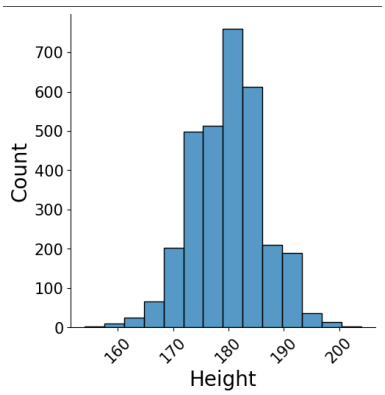
2 Data Understanding

Before starting the analysis, a preliminary evaluation of the two datasets was carried out, focusing on the syntactic and semantic correctness of their attributes. This first step aimed to identify and correct possible errors or inconsistencies in the data to ensure their reliability for subsequent research.

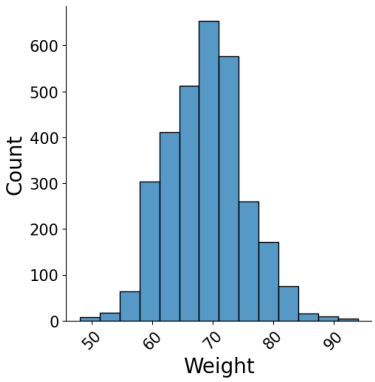
A thorough analysis was then performed on each dataset individually, followed by an analysis of the combined dataset, created by merging the two initial datasets. Each feature was carefully examined, with a particular focus on its distribution and correlations. Analyzing the distribution is essential as it helps to identify patterns, ensure that the values fall within expected ranges, and identify any irregularities. This step also highlights any potential skew or imbalance in the data, which may require adjustment or transformation. Additionally, examining the correlations provides insight into their relationships, helping to identify strong or weak associations that may influence further analysis.

2.1 Cyclist Understanding

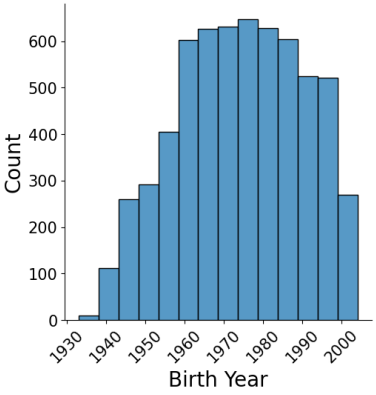
The years of birth in the dataset, as shown in Figure 1c, are predominantly from the last century, with values within the expected range. Cyclists born before 2000, especially those born in the 1960s, make up the largest groups. Regarding the `weight` and `height` attributes, according to data from external sources, the average height of cyclists is typically between 175 and 185 cm, while their weight is typically between 60 and 70 kg. The under-representation of values outside these ranges, as shown in Figures 1a and 1b, appears to be consistent with the natural statistical distribution of cycling, rather than a result of bias in the dataset.



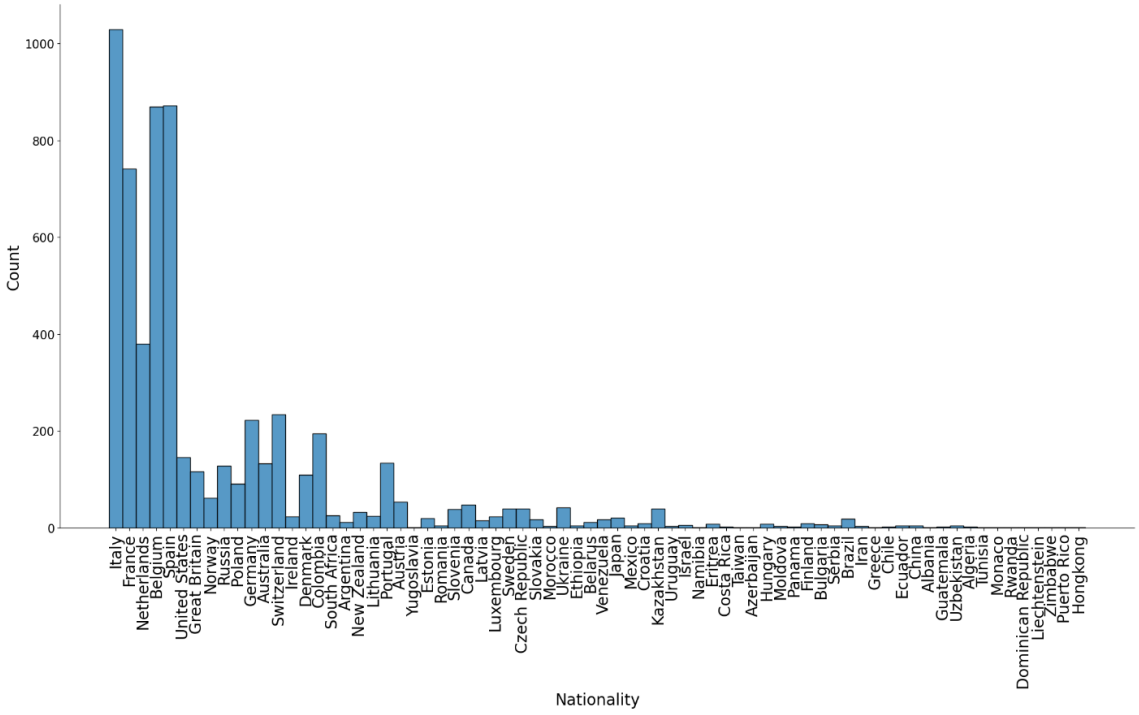
(a) height distribution



(b) weight distribution



(c) birth_year distribution



(d) nationality distribution.

Figure 1: Cyclist's feature distributions.

The dataset presents, as shown in figure 1d, that the majority of cyclists are of European nationality. There are plausible reasons for the prevalence of European cyclists in the dataset:

- Cycling has historically been a more prominent sport in Europe;
- Major cycling races are held in Europe, drawing athletes primarily from this region;
- Our dataset presents biases on the nationalities of selected cyclists or the races they participated in.

In terms of correlations within the dataset, the Pearson correlation between *birth_year* and *weight* and *height* is low (≤ 0.15), indicating a weak correlation. However, the correlation between *weight* and *height* is 0.72, indicating a strong positive relationship. This strong positive correlation between *weight* and *height* is to be expected as taller people tend to be heavier, and cyclists, by the nature of the sport, are generally not overweight.

2.2 Races Understanding

Starting from the name feature in the dataset, an analysis of race names revealed that different names were recorded for the same race. After conducting online researches to identify the official names, all variant occurrences were replaced

with their canonical counterparts. The term “ME" appearing in some names prompted further investigation, revealing that it refers to “Men Elite" and could potentially distinguish two different classes for the same race. Verification showed that “ME" and non-ME versions of the same competition never appear in the same year. Based on this finding, the “ME" nomenclature was determined to be another synonym for the canonical name of the race, and every occurrence was standardized accordingly. After this standardization, the dataset includes 32 distinct types of races. In the subsequent analysis of race participation, the bar chart in figure 2 shows the frequency of records for each different competition. Looking at this distribution, it can be seen that certain competitions, such as “O Gran Camino", “UAE Tour", and “Itzulia Basque Country", are less represented in the dataset in terms of participation, with these competitions accounting for less than or equal to 1% of the total records.

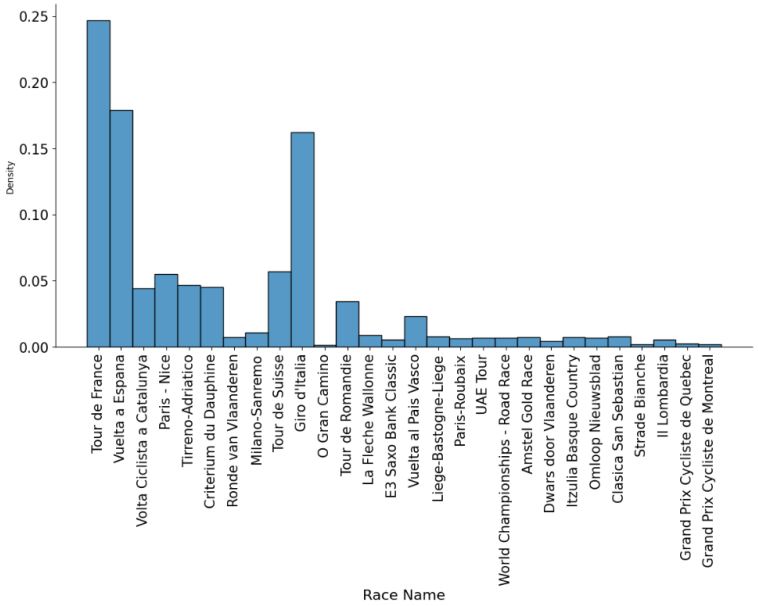


Figure 2: Races' participations.

The analysis shows that knowing the year and name of the competition alone does not provide a unique value of the `points` for each record in the dataset. To overcome this, the `_url` attribute is used, which contains information about both the competition and its stages. The investigation confirmed that each stage of a race has a unique score in the dataset, indicating that the score is not related to the race as a whole, but to the individual stages.

As for the `uci_points` attribute, it behaves in a similar way to the competition `points`, being specific to both the race and its stage. We observed that no `uci_points` records are found before 2001, which is consistent with research showing that this scoring system began in 2001.

A similar analysis of the `startlist_quality` values revealed its variation logic. It was found that for each race-edition pair, the `startlist_quality` is always unique, changing over the years for each race.

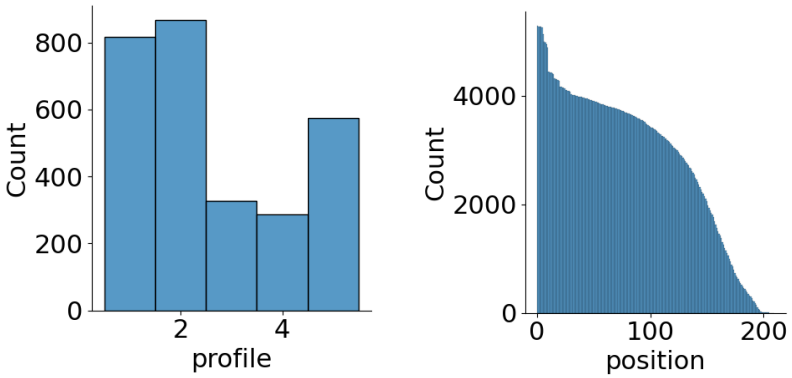
During the analysis of the `delta` feature, two main peculiarities were observed. Firstly, a small percentage of delta values are negative, which, according to the understanding, should be out of range and could indicate either an error or a rare situation, such as a cyclist retiring from the race due to injury or other similar events. Secondly, clusters of the same values appear for cyclists finishing in successive positions. While different delta values would be expected for each cyclist within the same race, whole groups of cyclists share the same delta. This is probably because cyclists cross the finish line in groups, and the delta reflects the time differences between these groups rather than individual times. In addition, 346 stages were found to have “suspicious" deltas, where cyclists in better positions have a worse time gap to the first-place athlete than those who finished after them.

The attributes `is_cobbled` and `is_gravel` each have only one value ($False$) along the whole dataset and therefore do not add any meaningful information to the analysis. Therefore, these two columns have been removed. In addition, the column `is_tarmac` has been renamed to `mostly_tarmac` to reflect the fact that “tarmac" terrain is always present, at least partially, on all tracks, with some races predominantly featuring this type of terrain (this information was confirmed by experts during the development).

An analysis of the `profile` attribute within the dataset reveals five distinct values, whose distribution is shown in figure 3a, which shows the predominance of flat and hilly stages.

Analyzing the distribution of the years associated with the records in the “races" dataset, it becomes evident from 4 that the majority of data are concentrated after approximately 1995. A noteworthy characteristic is the representation of the class of dates near 2020, where there is a clear absence of information compared to the surrounding years. This phenomenon is likely linked to the Covid-19 pandemic, which during that period had a significant impact, drastically reducing the opportunities for organizing and participating in public and sporting events such as cycling races.

Analysis of the `position` attribute provides insight into the ranking of cyclists within each race. The figure 3b illustrates the distribution of position values within the dataset, showing that most positions fall between 0 and 100, with those above 100 being less common. This is explainable by the fact that just some rarer races feature such high number of participants respect to the others.



(a) profile distribution. (b) position distribution.

Figure 3: profile and position distributions

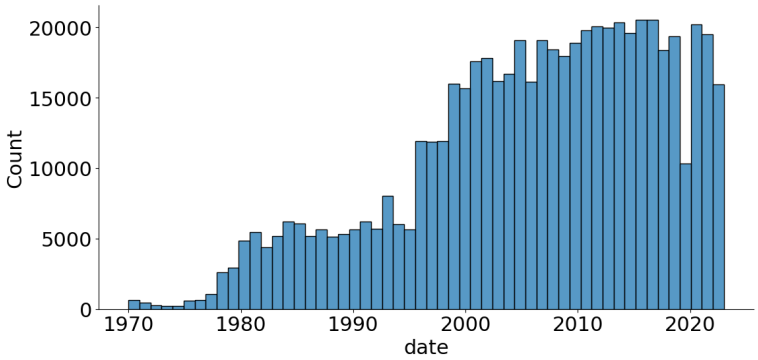


Figure 4: race_year distribution.

In general there is a lack of significant correlations between the different characteristics analyzed. The few correlations detected include a connection between `climb_total` and `profile`, where an increase in total elevation gain is often associated with a more challenging profile. Additionally, a strong correlation exists between `points` and `uci_points`, suggesting that the points awarded in the race generally align with the UCI's scoring criteria.

3 Data Preprocessing

The preprocessing phase involves cleaning the data, modifying and creating features, detecting and removing outliers, and applying PCA and UMAP for dimensionality reduction.

3.1 Data Cleaning

In this section, the focus is on the data cleaning process, which involved several key steps to ensure the dataset is ready for further analysis. The main steps in our cleaning process are the following:

- **Row Deletion**: Irrelevant or unnecessary rows are identified and removed to maintain the integrity and relevance of the dataset;
- **Imputation of Missing Values**: Missing data points are handled through imputation techniques, where values are filled based on various strategies to maintain completeness;
- **Correction of Erroneous Data**: Inconsistent or erroneous data points, such as negative values or outliers, are corrected to ensure data accuracy.
- **Type Conversion**: Following an initial review of the data, columns are converted to appropriate data types as necessary to facilitate subsequent analysis.

These steps were essential for ensuring the dataset is accurate, consistent, and complete, laying a solid foundation for further analysis and modeling.

3.1.1 Cyclist Cleaning

For the "cyclist" dataset, which consists of 6134 records, there are missing values in several attributes. Specifically, almost 49% records have missing data for `height`, same for `weight`, only 1 record for `nationality` and only 13 records for `birth_year`.

The missing values for `birth_year` were first manually imputed by checking available information online; for the remaining missing values, the mode was used. A similar approach was applied to the `nationality` feature. The only missing value was for the English cyclist Scott Davies.

Since `height` and `weight` are highly correlated, `weight` can be used to estimate `height`, and vice versa. Two methods were considered for imputing the missing values: one involves creating a small linear regression model, and the other uses the mean of the distributions divided into various bins. Both methods successfully impute the missing values. However while both methods produce similar results, the second method, based on the mean, maintains a distribution that is more consistent with the original data. Therefore, the imputation will be performed using the mean within the distribution. For the remaining missing values, only the records of cyclists without `height` and `weight`, and who were not present in any races, were deleted as they did not provide useful information.

3.1.2 Races Cleaning

For the "races" dataset, which consists of 589865 records, there are also missing values in several attributes. Specifically, 57% records have missing data for `uci_points`, almost 25% for `climb_total`, 25.1% for `profile`, almost 95% for `average_temperature`, 0.02% for `cyclist_age` and almost 27% for `cyclist_team`. Considering also that the same cyclist may appear multiple times (up to twice) in the same race ranking, the decision was made to delete the second occurrence and retain only the first.

Given the high proportion of missing data for `average_temperature`, it was assumed that it was not feasible to use this value effectively or to impute it accurately. Therefore, it was preferable to remove this attribute entirely.

The imputation of the feature `points` was performed by considering the scores from "nearby" races, meaning the score of the same race held in a different year, based on the year closest to the race that needed imputation. This approach was chosen because points typically do not change drastically over a few years. The same approach was used also for the `cyclist_team` feature.

The filling of `climb_total` and `profile` missing values followed a similar reasoning as the `weight` and `height` imputation, due to the high correlation between the former two features.

Considering the `delta` feature, negative deltas or values that are inconsistent with the cyclist's position are adjusted to maintain alignment with the race rankings. Starting from the first position, the records are checked, and if an invalid value is encountered, it is replaced with the previous one to ensure consistency throughout the race.

In the end, given the previously imputed `birth_year`, the `cyclist_age` was imputed exploiting the difference between the race year and the cyclist's birth year.

3.2 Feature Engineering

In the feature engineering phase, several new features were created, and existing ones were modified to improve the overall performance of the analysis. This process involved generating additional features to better capture the underlying patterns in the data, as well as transforming certain features through scaling or normalization. The following list shows the changes made to existing features and the new features created during the feature engineering process:

- Both the attributes `length` and `climb_total` represent distances in meters, but with values that typically reach or exceed kilometers. For this reason, these values were scaled by converting the unit of measure from meters to kilometers;
- The attribute `position` is normalised so that its values are more meaningful and comparable between different races in a way that is more invariant with respect to the total number of participants. After this normalization, each race edition had position values ranging from 0 (best position) to 1 (worst position);
- To extract as much information as possible, a new categorical attribute, `race_season`, was engineered. This decision was made particularly due to the fact that the original `temperature` attribute cannot be meaningfully exploited because of its large number of NaN values. The `race_season` attribute can serve as a useful proxy for similar types of information;
- The feature `cyclist_bmi` was created by combining the `height` and `weight` attributes, offering a more complete and descriptive measure of a cyclist's physical condition;
- The `cyclist_age_group` were created to balance the distribution of ages, ensuring a more even representation across different age ranges.
- The `cyclist_climb_power` function has been created to approximate the power/weight ratio by taking into account the difficulty of the race, the cyclist's performance and body composition through `weight`;
- The feature `race_physical_effort` was designed to summarise the technical difficulty of a given race by calculating its value based on the `length`, `climb_total` and `profile` of the track;
- The `race_prestige` attribute is designed to evaluate the relevance of a given race based on its participants and points value, incorporating the `startlist_quality` and `points` attributes;

- The `num_participants` feature was designed to be particularly useful for classification tasks, as a race with fewer participants is more likely to have a cyclist in the top-20;
- Three "mean" features were designed for the cyclists to be useful in the clustering process. These features are `mean_delta`, `mean_position` and `mean_climb_power`. Each of these features aggregates relevant information across races to provide a more generalised view of cyclists' performance and characteristics;
- Three "previous_mean" features were designed to represent the past performance of the cyclists, making them useful for both clustering and potentially classification. These features are `previous_mean_delta`, `previous_mean_position` and `previous_mean_climb_power`. Each of these features aggregates relevant information from previous races, providing a summarised view of the cyclist's past performance and characteristics;
- The final feature, `cyclist_previous_experience`, combines the `race_prestige` and `race_physical_effort` features from past races only, to serve as a proxy for the cyclist's experience.

3.3 Outlier Detection

Initially, several box plots were used to examine each feature and identify potential outliers. However, to remove outliers that could negatively affect the clustering and classification tasks, the Isolation Forest method was used.

The analyses reported in the following chapters will examine three types of clustering: clustering based solely on cyclists' data, clustering based solely on races' data, and clustering based on a combination of both cyclists' and races' data. To account for the unique context of each clustering type, three separate instances of Isolation Forest were applied, one for each case, to detect and remove outliers appropriately.

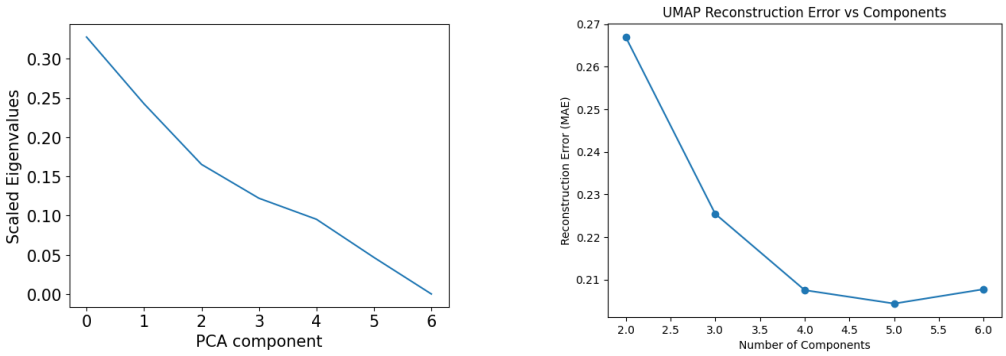
After applying this method, 8.8% of outliers were removed for the cyclists' data, 18.6% for the races' data, and 7.3% for the combined races and cyclists' data.

3.4 Feature Representation

For the feature representation analysis, $UMAP$ (Uniform Manifold Approximation and Projection) and PCA (Principal Component Analysis) were used to reduce the dimensionality of the dataset, which aided the clustering process. However, the components generated by these algorithms were not used directly for clustering in order to balance feature interpretability and dimensionality reduction. Instead, a qualitative 'visual inspection' was performed to identify which features of the dataset were more correlated with the components generated by $UMAP$ and PCA; for PCA, component selection was based on the "elbow" method, while for $UMAP$, the components that minimized the reconstruction error were selected. As this approach is a heuristic method based on visual judgment, for the three types of analysis, we removed features that did not align with any component of both algorithms.

3.4.1 Cyclist

The plot in Figure 5a does not display a clear elbow, indicating that there are no distinct principal components among the 6 analyzed, according to PCA.

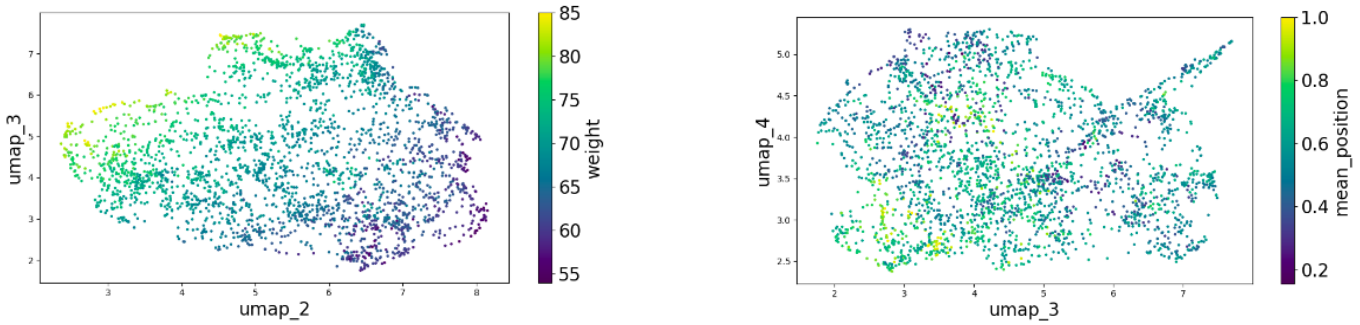


(a) PCA (b) UMAP

Figure 5: Plots for selection of components number for cyclists data

As mentioned earlier, the selection of the number of components for $UMAP$ was based on how well the reduced data could reconstruct the original data. As shown in Figure 5b, the minimum reconstruction error occurs at 5 components.

After analysing $UMAP$, it was found that only the `mean_position` attribute was not relevant. However, as PCA suggests that all attributes are important, it was decided to use all features for the clustering activities with the "cyclists" data (Figures 6a 6b).



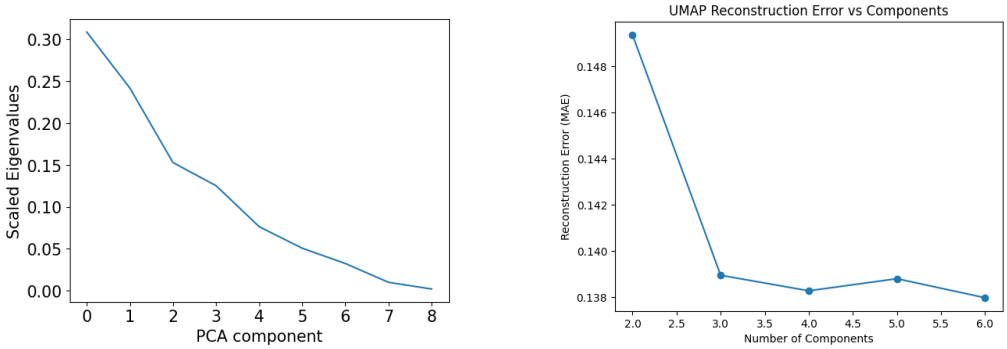
(a) UMAP good correlation example

(b) UMAP bad correlation example

Figure 6: Example of good and bad UMAP correlations for cyclists data

3.4.2 Races

For the “races" dataset, looking at the plot shown in Figure 7a, 4 principal components were selected for the PCA analysis. Considering all possible alignments of the features with the principal components identified by PCA, both the `startlist_quality` and `race_year` features generally show poor alignment.



(a) PCA

(b) UMAP

Figure 7: Plots for selection of components number for races data

In this $UMAP$ analysis, the selection is limited to a maximum of 6 components in order to keep the computations manageable. Within this range, the minimum number of components that provide a satisfactory representation of the data is considered, balancing the trade-off between computational feasibility and quality of feature representation.

The features `length`, `startlist_quality`, `race_year`, `race_prestige` and `num_participants` show weak correlations with the $UMAP$ components in several directions. Since both PCA and $UMAP$ are used for robustness in selecting the most representative features, and PCA in this particular case highlights `length`, `race_prestige` and `num_participants` as relevant, these features were chosen to be included in the clustering process.

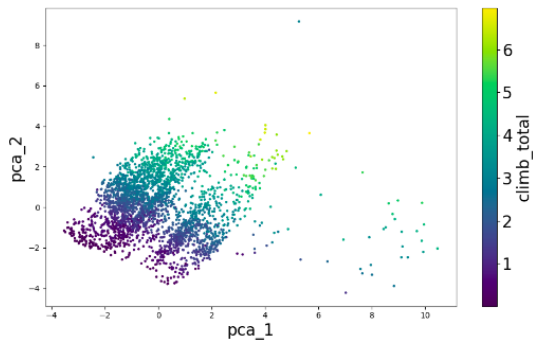
Therefore, the features that were decided not to be used for clustering activities with the “races" data are the attributes `startlist_quality` and `race_year`, as they show a weak correlation in both analyses. In addition, the `climb_percentage` feature is not used as almost all of its values are very low and similar to each other. This lack of variation could adversely affect the clustering process by over-compressing the points, potentially making them too similar and thus reducing the effectiveness of the clustering algorithm (Figures 8a 8b 8c 8d).

3.4.3 Ronde van Vlaanderen

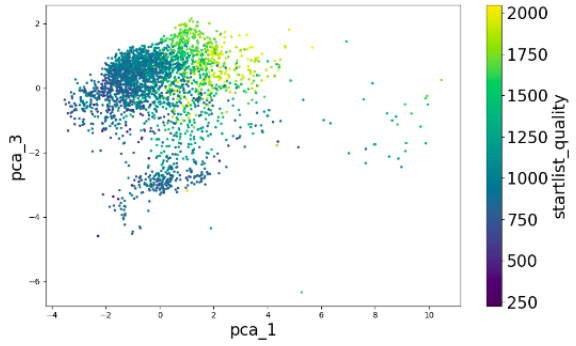
For the “Ronde van Vlaanderen" dataset, looking at the plot shown in Figure 9a, 6 principal components were selected for the PCA analysis. Considering all possible alignments of the features with the principal components identified by PCA, both the `startlist_quality`, `race_year`, `length`, `position`, `delta`, `race_prestige`, `num_participants` features generally show poor alignment.

As with the “races” dataset, the component selection is up to 6 components. According to the $UMAP$ results, almost all the features have a weak correlation with the components in several directions. This may be due to problems encountered in the process of selecting the number of components, which may have resulted in a potentially sub-optimal selection. Nonetheless, the PCA analysis was also performed, and therefore, the results from PCA are primarily considered.

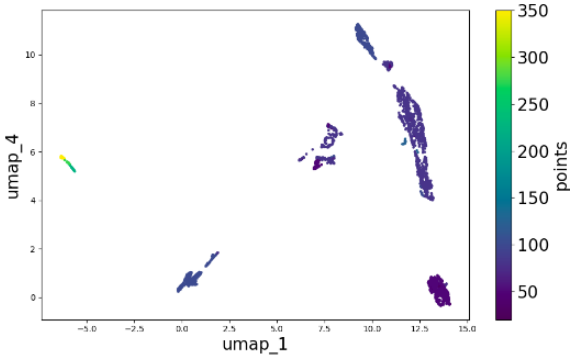
Therefore, the features that were decided not to be used for clustering activities with the “Ronde van Vlaanderen" data are the attributes `length`, `startlist_quality`, `position`, `delta`, `race_year`, `race_prestige` and `num_participants`, as they show a weak correlation in both analyses. In addition, the `cyclist_climb_power` and `previous_mean_cp` features are not used as almost all of the values are very low and similar to each other (Figures 10a 10b).



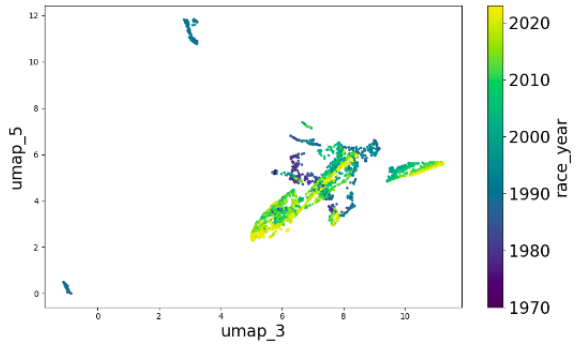
(a) PCA good correlation example



(b) PCA bad correlation example

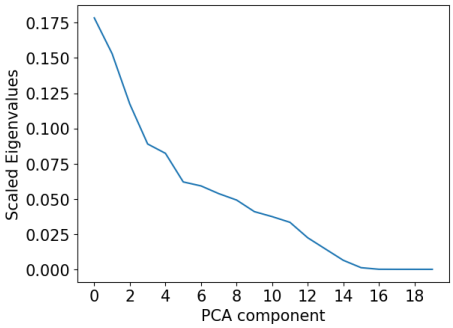


(c) UMAP good correlation example

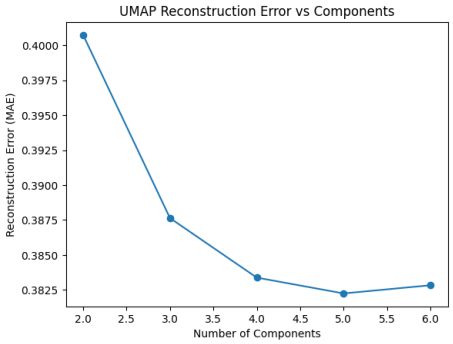


(d) UMAP bad correlation example

Figure 8: Example of good and bad PCA and UMAP correlations for races data

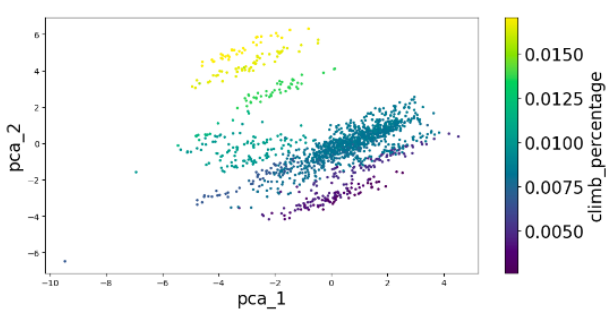


(a) PCA

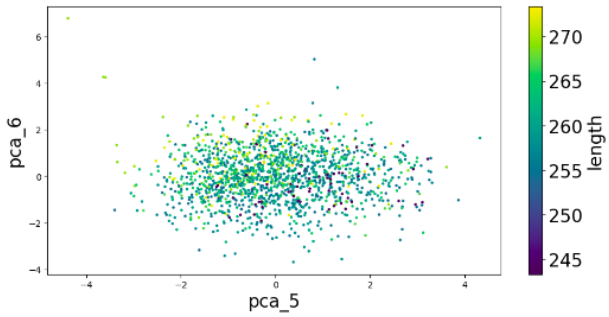


(b) UMAP

Figure 9: Plots for selection of components number for Rone van Vlaanderen race data



(a) PCA good correlation example



(b) PCA bad correlation example

Figure 10: Example of good and bad PCA correlations for the Ronde van Vlaanderen race data

4 Clustering

Clustering is a core data analysis technique for grouping similar data points. This document explores three widely used methods—DBSCAN, Hierarchical Clustering, and K-Means—applied to cycling and race-related datasets. Each method uncovers patterns and insights that enhance our understanding of cyclist performance and race characteristics.

DBSCAN was employed for its ability to handle clusters of varying density and outliers, by tuning parameters like `eps` and `min_samples` with a parameter selection process.

Hierarchical Clustering revealed nested relationships through dendrograms using linkage methods such as Ward and Average, with Ward linkage often yielding the most interpretable results.

K-Means, a simpler method, partitioned data into predefined clusters, emphasizing features like climb power, weight, and race length. Optimal cluster counts were determined using the Elbow Method and Silhouette Score.

Despite the insights, challenges like dominant clusters and overlapping groups suggest the need for refined features or alternative approaches. This analysis demonstrates the potential of clustering in exploring complex datasets while acknowledging its limitations.

4.1 Cyclist - Density-Based Clustering (DBSCAN)

This analysis employed the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to cluster cyclist data. DBSCAN was chosen for its capacity to identify clusters of varying densities and its robustness to outliers, a common occurrence in real-world datasets. Seven features were selected based on prior feature engineering and dimensionality reduction using PCA and UMAP: `mean_cyclist_cp`, `mean_delta`, `mean_position`, `birth_year`, `height`, `weight`, and `cyclist_bmi`. These features were found to align closely with the principal components.

4.1.1 Model Tuning and Analysis

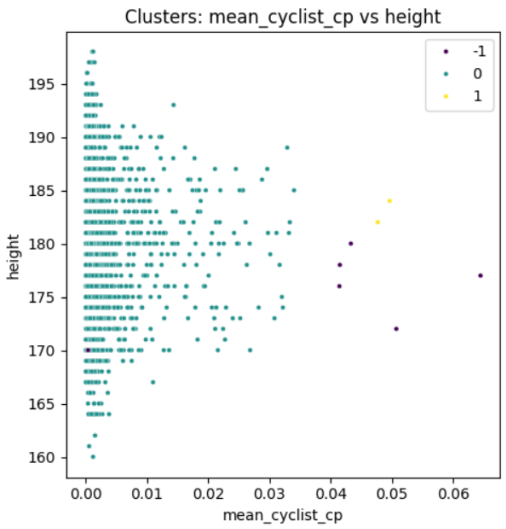
The analysis focused on tuning DBSCAN's key parameters: `eps` (maximum neighbor distance) and `min_samples` (minimum points to form a cluster). The tuning process followed these steps:

1. **Calculate k-distances:** For each data point, compute the distance to its k-th nearest neighbor, where $k =$ `min_samples` $- 1$.
2. **Sort k-distances:** Arrange the distances in ascending order.
3. **Identify the knee-point:** Use the `KneeLocator` library to locate the maximum curvature in the k-distance plot, corresponding to the best `eps` for that values of `min_samples`, marking the transition between dense and sparse regions.

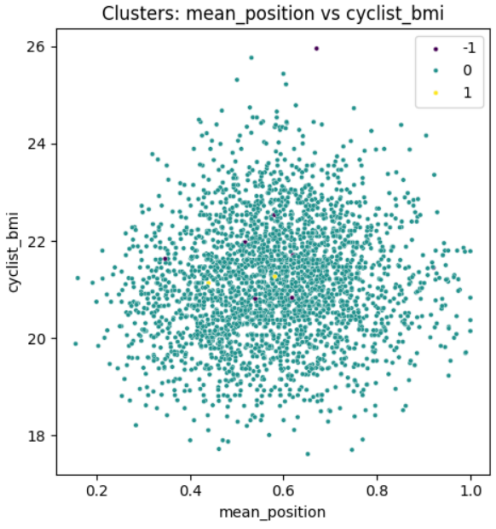
Results showed that setting `min_samples = 2` yielded three clusters with a higher silhouette score than `min_samples = 3`, which consistently identified the same number of clusters. These findings underscore the sensitivity of DBSCAN to parameter adjustments.

4.1.2 Observations

Most records were concentrated in one large cluster, with another significant cluster containing a minority of points as shown in Figure 11b. Outliers were labeled as -1.



(a) Clustering respect to the features `height` and `mean_cyclist_cp`



(b) Another perspective on the same clustering result. Here we can clearly observe the large main cluster

Figure 11: Plots showing the main results of DBSCAN applied to cyclist data.

The feature `mean_cyclist_cp` emerged as a key differentiator, separating cyclists with a mean climb power above 0.04 into distinct performance groups (Figure 11a).

The dominance of a single large cluster limits interpretability, suggesting uniform data density or that the chosen features may not fully capture finer variations. The silhouette score, while informative, might inadequately evaluate clustering quality in imbalanced or overlapping datasets. Future analysis using alternative metrics or clustering techniques could improve the robustness of these insights.

4.2 Races - Density-Based Clustering (DBSCAN)

This section applies DBSCAN to races data, leveraging its ability to handle variable densities and detect outliers. Key features were identified through PCA and UMAP: `points`, `length`, `climb_total`, `race_physical_effort`, `race_prestige`, and `num_participants`.

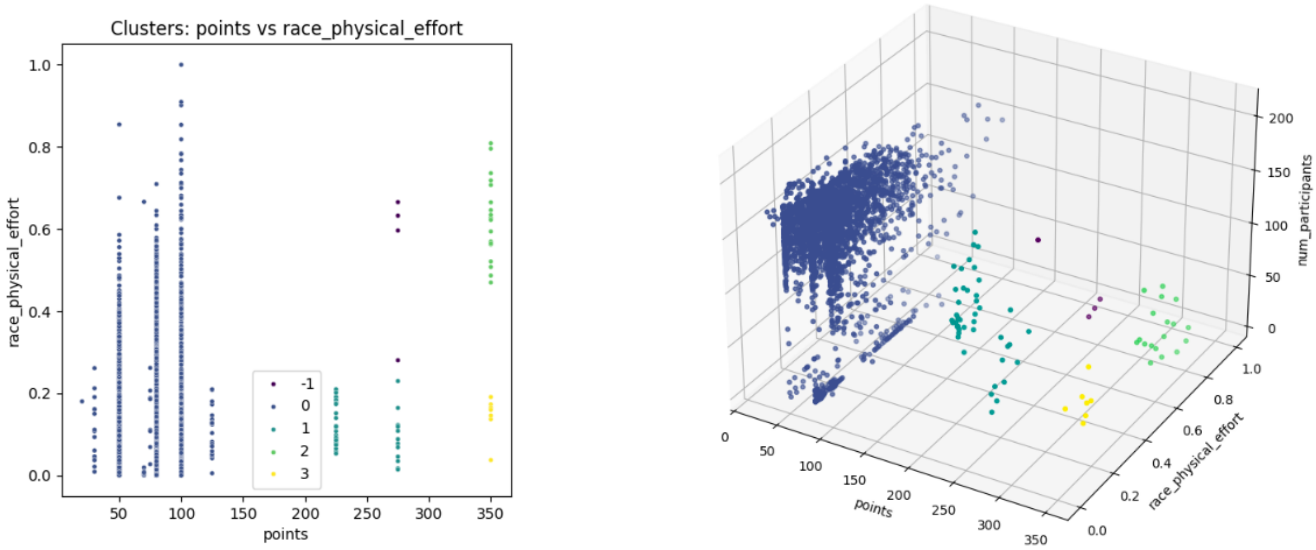
4.2.1 Model Tuning and Analysis

DBSCAN parameters were optimized to balance the number of clusters and the silhouette score. With `min_samples = 7`, five clusters were identified, achieving the highest silhouette score among configurations.

4.2.2 Observations

The clustering revealed a dominant large cluster (Figure 12b), with additional clusters distinguished by `race_physical_effort` and `points`. Notable patterns include:

- Races grouped by `points` into three categories: 0–150, 150–300, and over 300 (Figure 12a).
- Among races offering over 300 points, subgroups emerged based on `race_physical_effort` exceeding or below 0.3 (Figure 12a).



(a) Clustering respect to the features `race_physical_effort` and `points`

(b) Another perspective on the same clustering result. Here we can clearly observe the large main cluster

Figure 12: Plots showing the main results of DBSCAN applied to races data.

While the silhouette score exceeded 0.55, indicating a reasonable configuration, it was modest, suggesting room for improvement.

The clustering's dominance by a single large cluster and a moderate silhouette score reflect challenges in the dataset's density distribution or feature differentiation. Complementary metrics or alternative techniques may yield more robust insights.

4.3 Ronde van Vlaanderen Results - Density-Based Clustering (DBSCAN)

This analysis applied DBSCAN to cyclist results from the Ronde van Vlaanderen race, using 11 features selected through PCA and UMAP: `birth_year`, `weight`, `height`, `climb_total`, `cyclist_age`, `cyclist_bmi`, `climb_percentage`, `race_physical_effort`, `previous_mean_position`, `previous_mean_delta`, and `cyclist_previous_experience`.

4.3.1 Model Tuning and Analysis

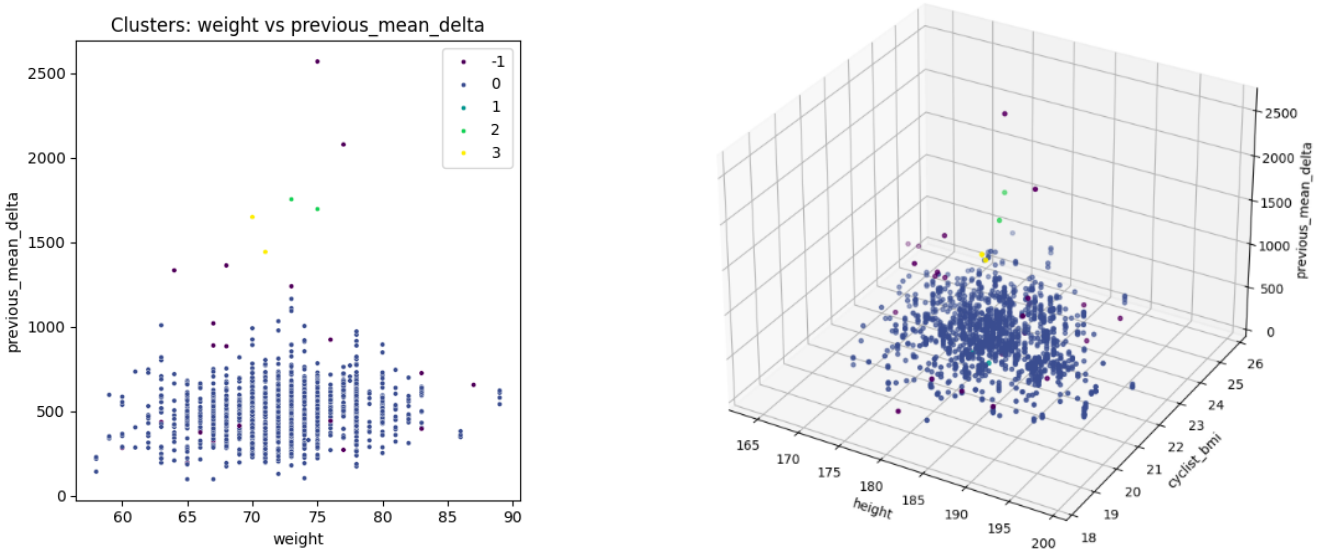
The parameters `eps` and `min_samples` were tuned to balance cluster count and silhouette score, with `min_samples = 2` yielding the best configuration.

4.3.2 Observations

The clustering revealed a dominant large cluster (Figure 13b, with additional clusters distinguished by `race_physical_effort` and `points`. Notable patterns include:

- Races grouped by `points` into three categories: 0–150, 150–300, and over 300 (Figure 13a).

- Among races offering over 300 points, subgroups emerged based on `race_physical_effort` exceeding or below 0.3 (Figure 13a).



(a) 2D plot

(b) 3D plot

Figure 13: Plots showing the main results of DBSCAN applied to the Ronde van Vlaanderen race data.

2D and 3D visualizations showed weak correlations between features, such as `previous_mean_delta`, and the clusters. However, the small dataset size and sparse cluster distribution limited interpretability.

4.4 Cyclists - Hierarchical Clustering

This analysis applied the Hierarchical Clustering algorithm to cyclists' data to identify nested groupings and visualize clustering using dendrograms. Seven features were used: `mean_cyclist_cp` (climb power), `mean_delta`, `mean_position`, `birth_year`, `height`, `weight`, and `cyclist_bmi`. These features were selected based on their relevance from dimensionality reduction (UMAP, PCA) and their descriptive power for the dataset.

4.4.1 Detailed Analysis

Hierarchical Clustering was performed using four linkage methods: Complete, Single, Average, and Ward, with results interpreted through dendrograms and cluster features.

4.4.1.1 Complete Linkage The dendrogram (Figure 14a) showed a heavily unbalanced hierarchy, driven mainly by `mean_cyclist_cp`. Cyclists with high climb power were grouped into a yellow cluster (Figure 15a).

4.4.1.2 Single Linkage The Single method produced an even less balanced dendrogram (Figure 14b), resulting in clusters that failed to meaningfully separate data points (Figure 15b).

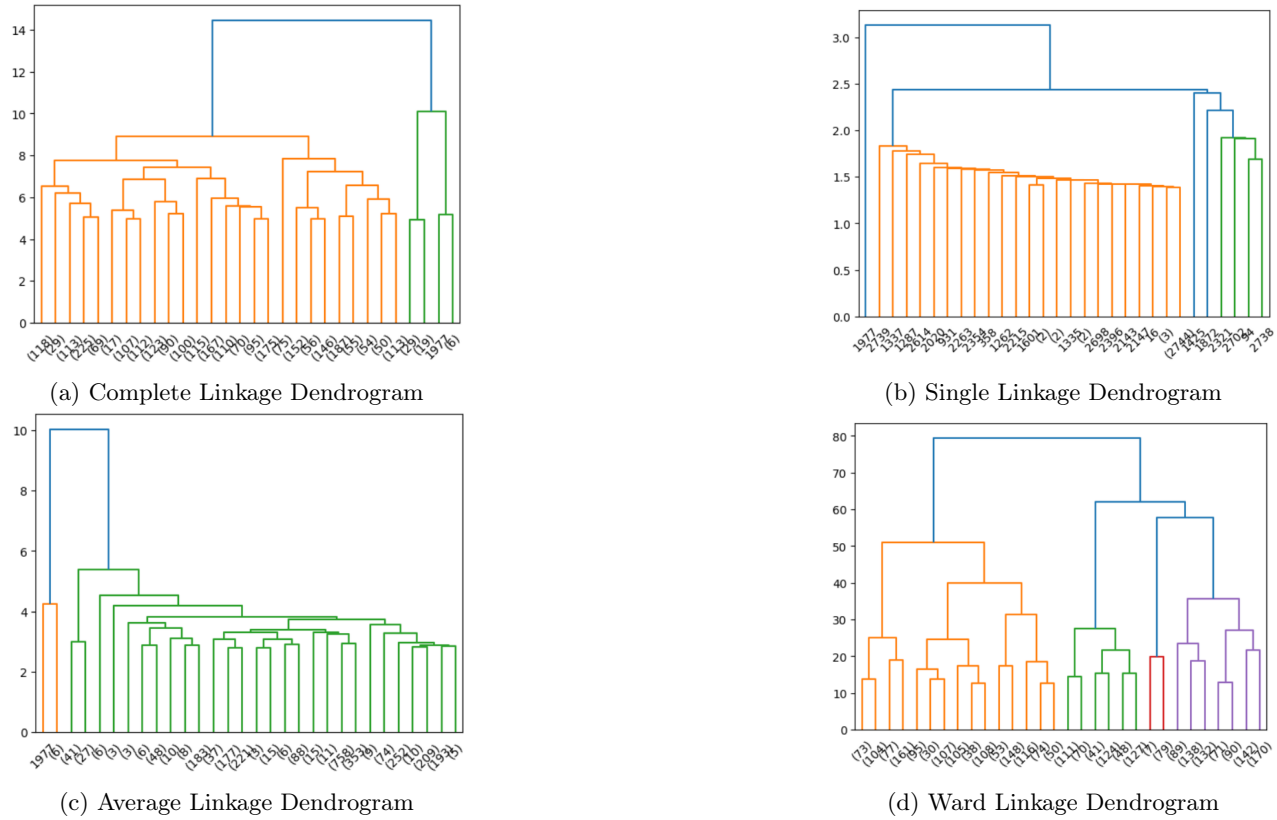
4.4.1.3 Average Linkage The Average method also created an unbalanced hierarchy (Figure 14c) dominated by `mean_cyclist_cp` (Figure 15c).

- **Cluster 1 (purple)**: High `mean_cyclist_cp`.
- **Cluster 2 (green)**: Medium `mean_cyclist_cp`.
- **Cluster 3 (yellow)**: Low `mean_cyclist_cp`.

4.4.1.4 Ward Linkage The Ward method yielded a more balanced hierarchy (Figure 14d). Cluster characteristics (Figure 15d) were identified using 2D and 3D visualizations :

- **Cluster 1 (purple)**: High `birth_year`, medium-high `mean_delta`.
- **Cluster 2 (light blue)**: Medium-low `birth_year`, medium-high `weight`.
- **Cluster 3 (yellow)**: Low `mean_delta`, `mean_position`, and `weight`.

However, Clusters 1 and 3 overlapped significantly, suggesting unresolved complexities or interdependencies.



(a) Complete Linkage Dendrogram

(b) Single Linkage Dendrogram

(c) Average Linkage Dendrogram

(d) Ward Linkage Dendrogram

Figure 14: Dendrograms showing the balancing of the results of Hierarchical Clustering applied to cyclists data.

4.4.2 Observations

The linkage methods produced varying outcomes:

- Complete, Single, and Average methods yielded unbalanced clusters primarily influenced by `mean_cyclist_cp`.
- The Ward method produced more balanced clusters and highlighted additional features (`birth_year`, `mean_delta`, `mean_position`, `weight`) contributing to separation.

A key limitation was the unbalanced hierarchies across most methods. The Ward method, while more balanced, showed overlap between Clusters 1 and 3, indicating unresolved data patterns. Future work could explore alternative clustering techniques or incorporate additional features to better capture data nuances.

4.5 Races - Hierarchical Clustering

This analysis applied Hierarchical Clustering to races data using six features: `points`, `length`, `climb_total`, `race_physical_effort`, `race_prestige`, and `num_participants`. These features were identified through data preprocessing and dimensionality reduction (UMAP, PCA) as most aligned with the dataset's principal components.

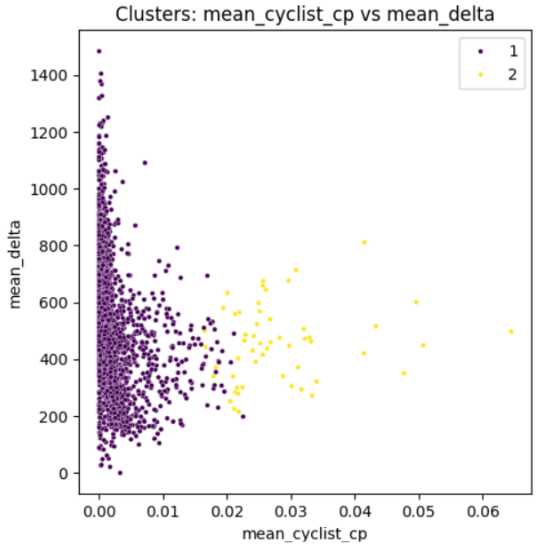
4.5.1 Detailed Analysis

Clustering was performed using four linkage methods: Complete, Single, Average, and Ward, with dendrograms and feature values examined to interpret cluster characteristics.

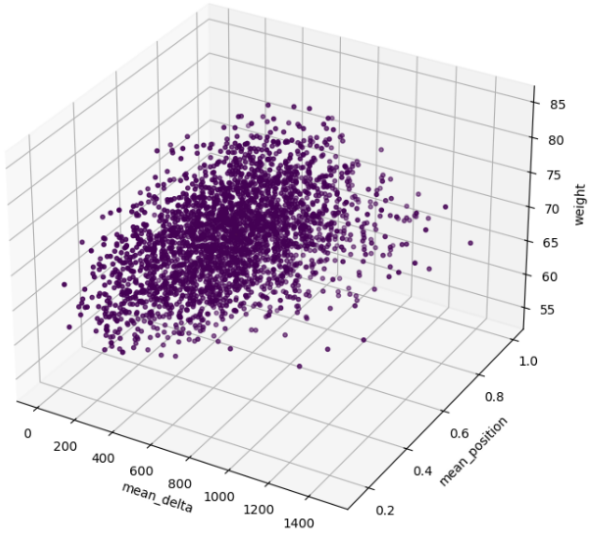
4.5.1.1 Complete Linkage The Complete method produced an imbalanced dendrogram (Figure 16a). Clusters were characterized as follows (Figure 17a):

- **Cluster 1 (purple)**: Medium-high `points`.
- **Cluster 2 (blue)**: High `race_prestige` and `points`.
- **Cluster 3 (light blue)**: Low `race_prestige` and `points`, medium `length`, high `race_physical_effort`.
- **Cluster 4 (green)**: Low `length` and `race_prestige`.
- **Cluster 5 (yellow)**: Low `race_prestige` and `points`, medium `length`, medium-low `race_physical_effort`.

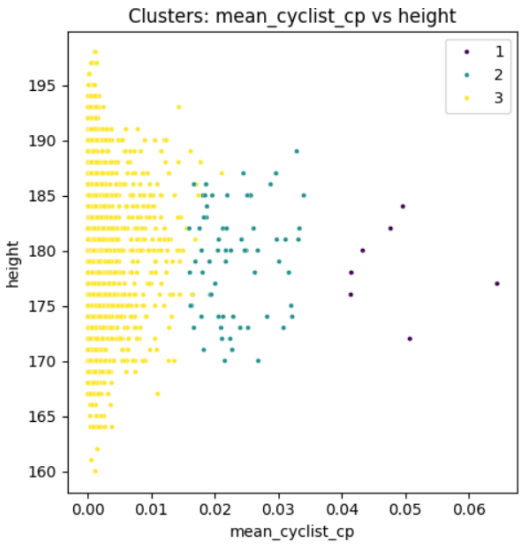
`Points` was pivotal for distinguishing Clusters 1 and 2, while features like `race_physical_effort` were needed to separate Clusters 3 and 5. The attribute `length` has been essential to distinguish between Clusters 4 and 5.



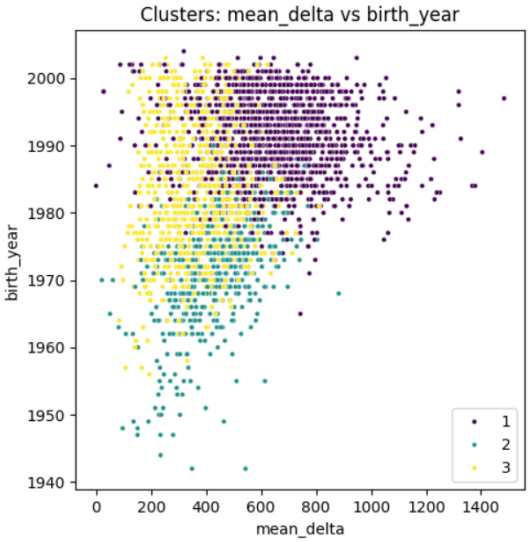
(a) Complete Linkage Results



(b) Single Linkage Results



(c) Average Linkage Results



(d) Ward Linkage Results

Figure 15: Plots showing the main results of Hierarchical Clustering applied to cyclists data.

4.5.1.2 Single Linkage The Single method yielded the following dendogram (Figure 16b) and clusters (Figure 17b):

- **Cluster 1 (purple)**: Low `points`.
- **Cluster 2 (blue)**: Medium-high `points`, low `race_physical_effort`.
- **Cluster 3 (light blue)**: High `points` and `race_prestige`.
- **Cluster 4 (green)**: Medium-high `points`, high `race_physical_effort`.
- **Cluster 5 (yellow)**: Medium-high `points`, medium `race_physical_effort`, high `race_prestige`.

`Points` remained a key feature, with `race_prestige` providing additional differentiation. However, Cluster 5 (yellow) contained only one data point, reducing its interpretability.

4.5.1.3 Average Linkage The Average method, whose dendogram is shown in Figure 16c, provided more interpretable clusters (Figure 17c):

- **Cluster 1 (purple)**: High `points` and `race_prestige`.
- **Cluster 2 (light blue)**: Low `points` and `race_prestige`.
- **Cluster 3 (yellow)**: Medium `points`.

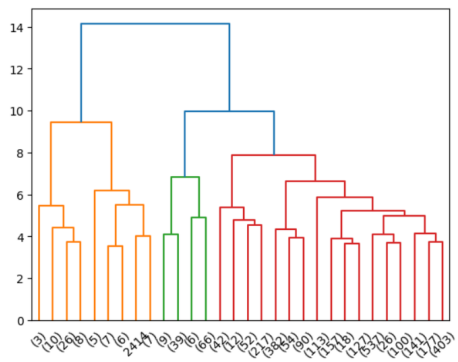
While primarily driven by `points`, this method offered cleaner separations than Complete or Single methods.

4.5.1.4 Ward Linkage

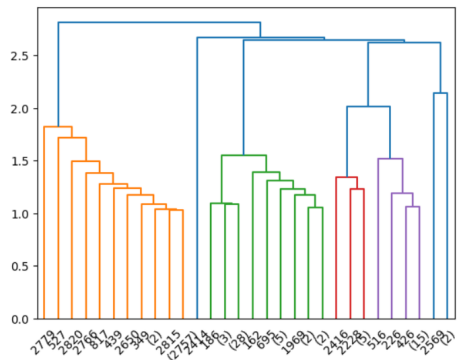
Ward linkage, dendogram shown in Figure 16d, yielded the following clusters (Figure 17d):

- **Cluster 1 (purple)**: Low `length`, medium-high `num_participants`.
- **Cluster 2 (blue)**: Medium `length`, medium-high `num_participants`, medium-low `climb_total`.
- **Cluster 3 (light blue)**: High `points`.
- **Cluster 4 (green)**: Medium `length`, medium-high `num_participants`, medium-high `climb_total`.
- **Cluster 5 (yellow)**: Low `num_participants`.

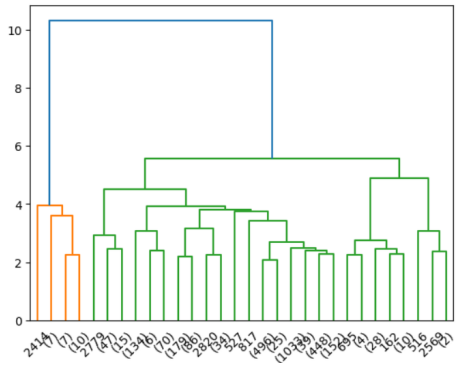
Clusters were distinguished by a combination of `points`, `length`, `num_participants`, and `climb_total`, with `points` being critical for Cluster 3.



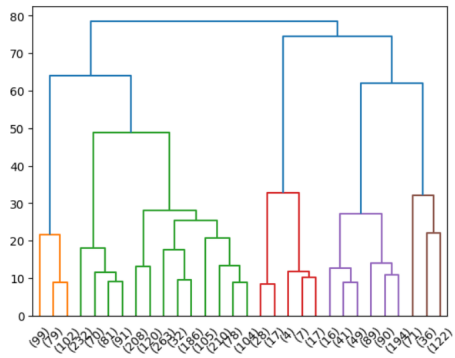
(a) Complete Linkage Dendrogram



(b) Single Linkage Dendrogram



(c) Average Linkage Dendrogram



(d) Ward Linkage Dendrogram

Figure 16: Dendrograms showing the balancing of the results of Hierarchical Clustering applied to races data.

4.5.2 Observations

Key findings included:

- Complete, Single, and Average methods relied heavily on `points` and `race_prestige` for separation.
- Ward linkage utilized additional features (`length`, `num_participants`, `climb_total`) for improved cluster differentiation.

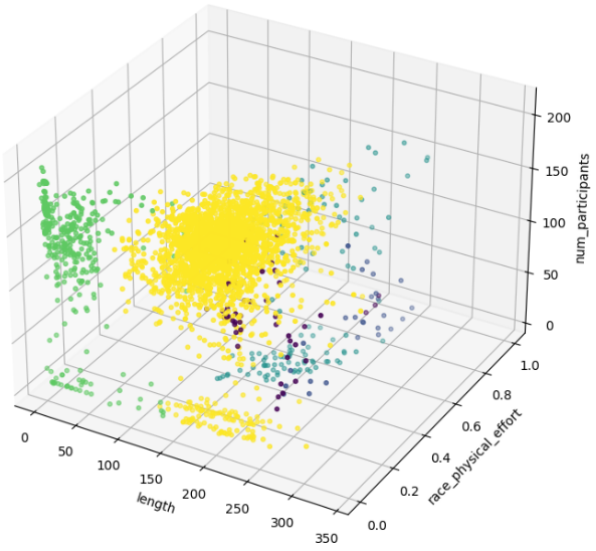
A limitation was the overlap between certain clusters, highlighting the need for alternative features or methods to better separate groups.

4.6 Ronde van Vlaanderen Results - Hierarchical Clustering

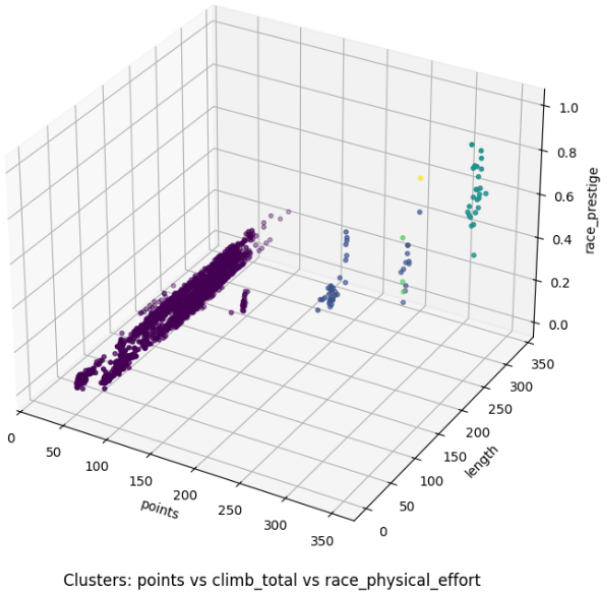
This analysis applied Hierarchical Clustering to results data from the Ronde van Vlaanderen race, using 11 features: `birth_year`, `weight`, `height`, `climb_total`, `cyclist_age`, `cyclist_bmi`, `climb_percentage`, `race_physical_effort`, `previous_mean_position`, `previous_mean_delta`, and `cyclist_previous_experience`. These features were identified through UMAP and PCA, following rigorous data cleaning and feature engineering.

4.6.1 Detailed Analysis

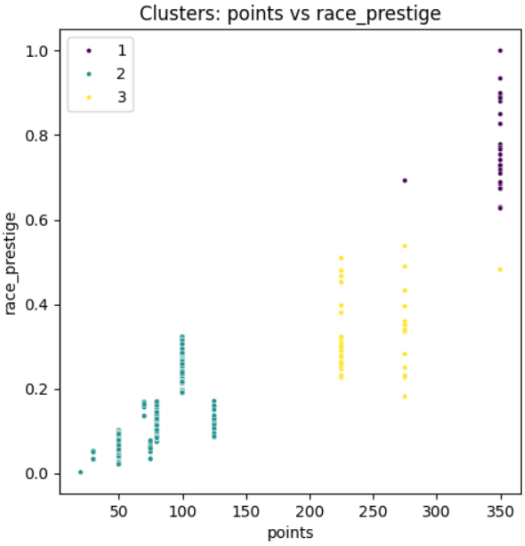
Hierarchical Clustering was performed using four linkage methods: Complete, Single, Average, and Ward. Dendrograms (Figures 18a, 18b, 18c, 18d) and feature characteristics were analyzed for cluster interpretation.



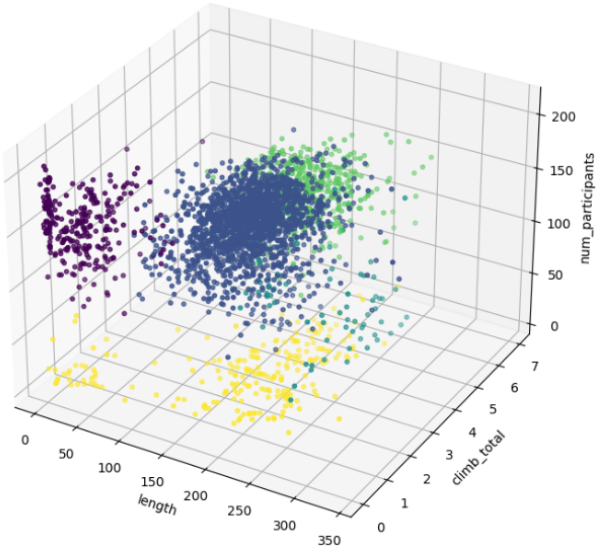
(a) Complete Linkage Results



(b) Single Linkage Results



(c) Average Linkage Results



(d) Ward Linkage Results

Figure 17: Plots showing the main results of Hierarchical Clustering applied to races data.

4.6.1.1 Complete Linkage

The Complete method produced the following clusters (Figure 19a):

- **Cluster 1 (purple)**: High `previous_mean_delta`.
- **Cluster 2 (blue)**: High `climb_total` and `climb_percentage`, low `previous_mean_delta`.
- **Clusters 3 (green) and 4 (yellow)**: Medium-low `climb_total` and `climb_percentage`, low `previous_mean_delta`. These clusters lacked clear boundaries.

Clusters 3 and 4 likely require higher-dimensional analysis for clearer differentiation.

4.6.1.2 Single Linkage

Single Linkage resulted in one large cluster with a few outliers (Figure 19b). The limited separation provided by this method made it unsuitable for meaningful subgroup identification.

4.6.1.3 Average Linkage

The Average method offered a more interpretable structure (Figure 19c):

- **Cluster 1 (purple)**: High `previous_mean_delta`.
- **Cluster 2 (light blue)**: Low `previous_mean_delta`, very high `race_physical_effort`, `climb_total`, and `climb_percentage`.
- **Cluster 3 (yellow)**: Low `previous_mean_delta`, with `race_physical_effort`, `climb_total`, and `climb_percentage` ranging from low to high.

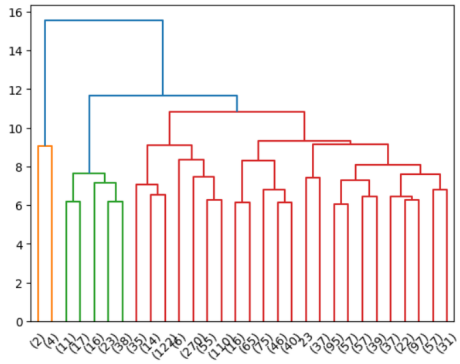
`Previous_mean_delta` was the primary driver of separation, distinguishing Cluster 1 (purple) from others. Clusters 2 and 3 were differentiated by the extreme values in Cluster 2 compared to the variable ranges in Cluster 3.

4.6.1.4 Ward Linkage

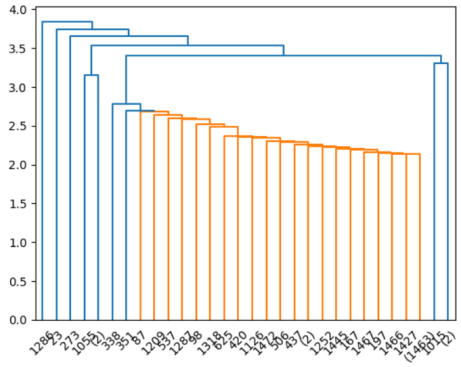
Ward linkage produced the following clusters (Figure 19d):

- **Cluster 1 (purple)**: High `previous_mean_delta`, `climb_total`, and `climb_percentage`.
- **Clusters 2 (blue) and 3 (yellow)**: Medium-low `climb_total`, `climb_percentage`, and `previous_mean_delta`, with insufficient feature separation.

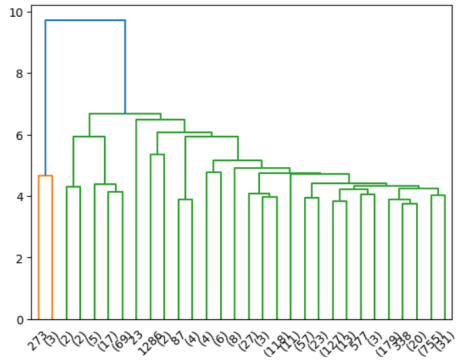
Clusters 2 and 3, like in Complete Linkage, require higher-dimensional analysis for clearer differentiation.



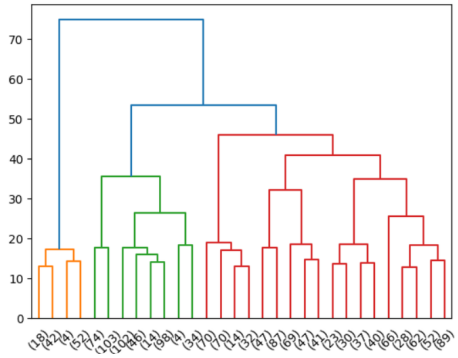
(a) Complete Linkage Dendrogram



(b) Single Linkage Dendrogram



(c) Average Linkage Dendrogram



(d) Ward Linkage Dendrogram

Figure 18: Dendrograms showing the balancing of the results of Hierarchical Clustering applied to Ronde van Vlaanderen race data.

4.6.2 Observations

Key findings include:

- `Previous_mean_delta` was the primary feature driving cluster separation across all methods.
- Features such as `climb_total`, `climb_percentage`, and `race_physical_effort` significantly contributed to distinguishing extreme cases.
- Single Linkage was ineffective, while Average and Ward methods provided more interpretable clusters.
- Overlaps in certain clusters (e.g., clusters 2 and 3 in Complete and Ward methods) indicate a need for higher-dimensional analyses or alternative features for better separation.

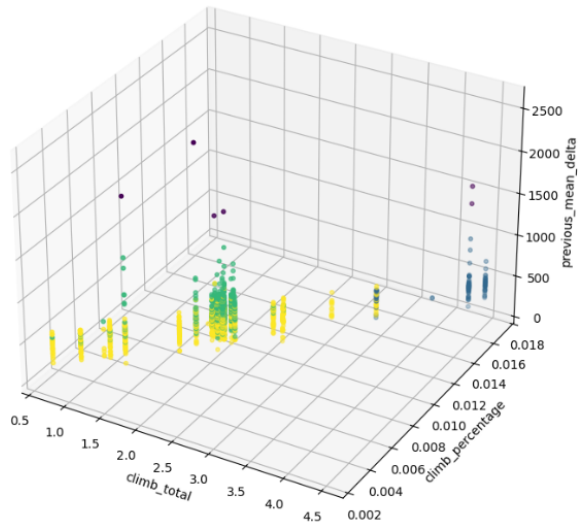
Future work could explore alternative clustering techniques or additional features to address these overlaps and improve cluster interpretability.

4.7 Cyclists - K-Means Clustering

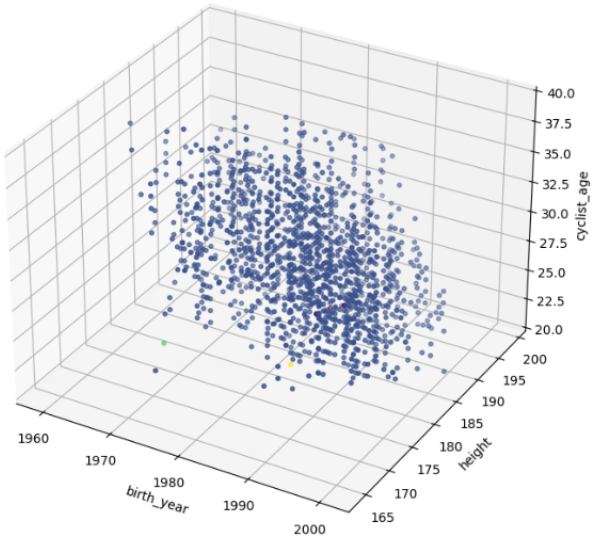
This analysis applied the K-Means clustering algorithm to cyclists' data using 7 features: `mean_cyclist_cp` (climb power), `mean_delta`, `mean_position`, `birth_year`, `height`, `weight`, and `cyclist_bmi`. These features were selected through a comprehensive data understanding phase and verified for relevance using UMAP and PCA analyses.

4.7.1 Cluster Analysis

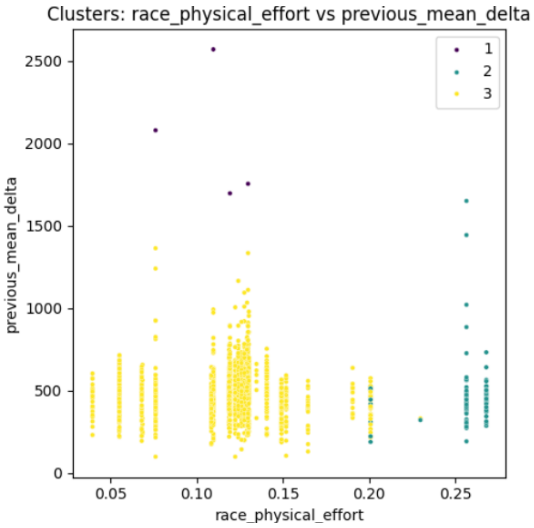
The optimal number of clusters was determined using the Elbow Method and Silhouette Score. Based on this evaluation, 4 clusters were chosen, balancing compactness and separation for meaningful interpretation.



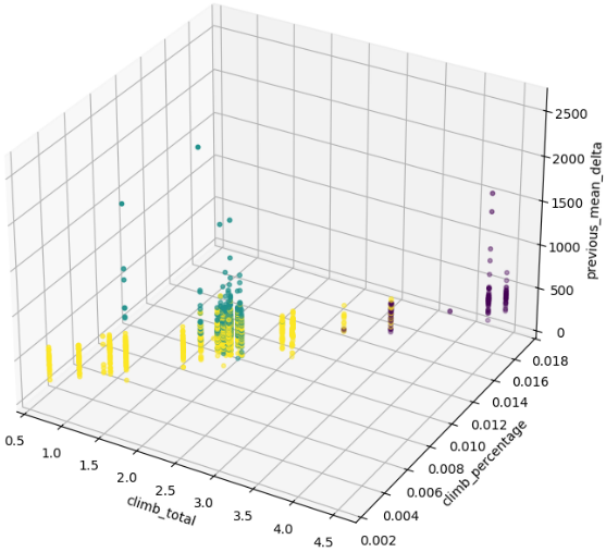
(a) Complete Linkage Results



(b) Single Linkage Results



(c) Average Linkage Results



(d) Ward Linkage Results

Figure 19: Plots showing the main results of Hierarchical Clustering applied to Ronde van Vlaanderen race data.

K-Means clustering results were analyzed through 2D and 3D visualizations. Cluster centroids highlighted key group characteristics, with `mean_cyclist_cp`, `weight`, and `mean_delta` emerging as the most significant features for distinguishing clusters.

The clusters were identified as follows:

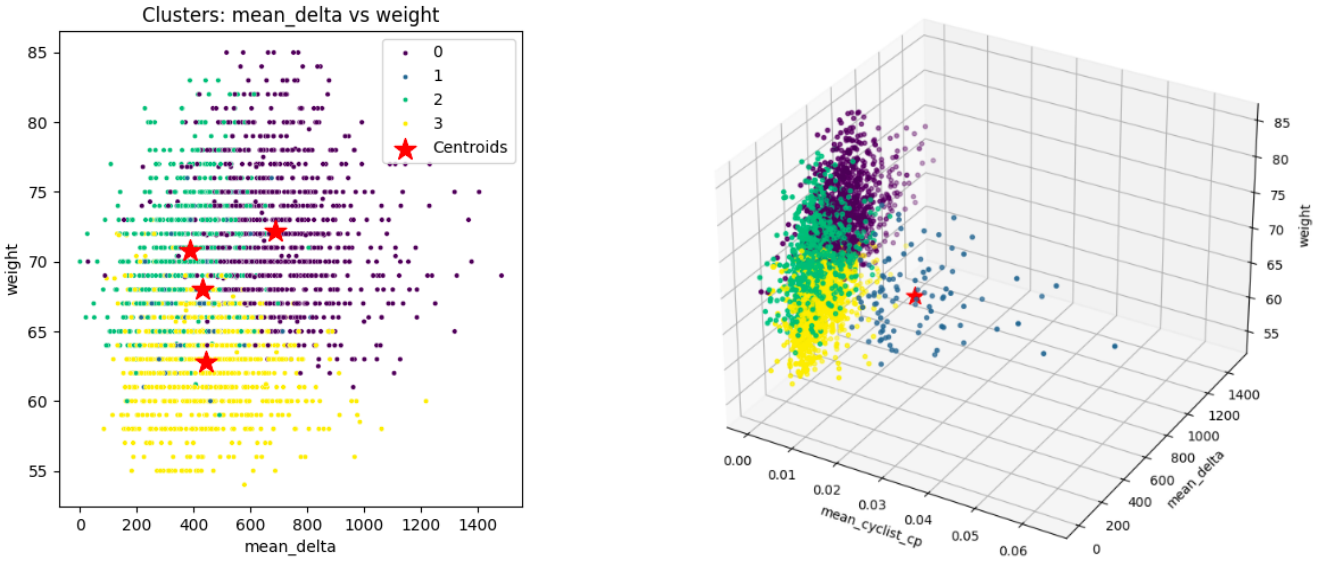
- **Cluster 0 (purple)**: Medium-high `mean_delta`, medium-high `weight`, and low `mean_cyclist_cp`.
- **Cluster 1 (blue)**: Medium-high `mean_cyclist_cp`.
- **Cluster 2 (green)**: Low `mean_delta`, medium-high `weight`, and low `mean_cyclist_cp`.
- **Cluster 3 (yellow)**: Medium-low `mean_delta`, low `weight`, and low `mean_cyclist_cp`.

4.7.2 Observations

This analysis demonstrates that `mean_cyclist_cp` is pivotal for outlier identification (Cluster 1, Figure 20b), while `mean_delta` and `weight` are key for separating the remaining clusters (Figure 20a). These results provide a meaningful segmentation of cyclists, offering valuable insights into their physical and performance profiles.

4.8 Races - K-Means Clustering

This study applied the K-Means clustering algorithm to race-related data, utilizing 6 features: `points`, `length`, `climb_total`, `race_physical_effort`, `race_prestige`, and `num_participants`. These features were chosen following rigorous data cleaning and feature engineering phases. UMAP and PCA analyses confirmed their strong alignment with the principal components, validating their suitability for clustering.



(a) 2D plot

(b) 3D plot

Figure 20: Plots showing the main results of K-Means applied to the Cyclists data.

4.8.1 Cluster Analysis

The optimal number of clusters was determined using the Elbow Method and Silhouette Score.

Based on this analysis, a configuration of 6 clusters was selected, offering a balance between compactness and separation for meaningful interpretation.

Cluster results were explored through 2D and 3D visualizations. Analysis of feature values and cluster centroids revealed the distinguishing characteristics of each group.

The identified clusters were:

- **Cluster 0 (purple)**: Medium-low `race_prestige`, moderate `climb_total`, medium `length`, and high `num_participants`.
- **Cluster 1 (blue)**: Low `race_prestige`, low `num_participants`, and medium `length`.
- **Cluster 2 (light blue)**: Low `length`, `climb_total`, and `race_prestige`.
- **Cluster 3 (aqua green)**: Medium-low `race_prestige`, moderate `climb_total`, medium `length`, and high `num_participants`.
- **Cluster 4 (green)**: Medium-low `race_prestige`, medium-high `climb_total`, medium `length`, and high `num_participants`.
- **Cluster 5 (yellow)**: High `points`, `length`, and `race_prestige`.

4.8.2 Observations

This K-Means clustering analysis provided a meaningful segmentation of races, highlighting unique patterns and offering insights into the factors that differentiate races based on key features such as `race_prestige`, `length`, and `num_participants` (Figures 21a 21b).

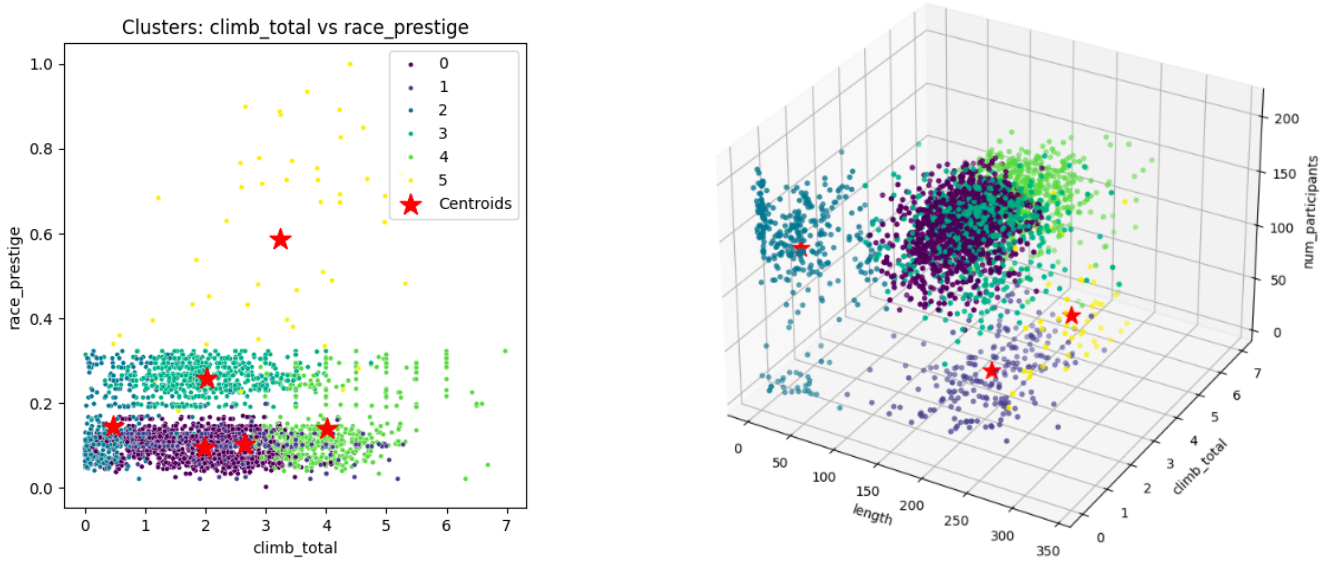
4.9 Ronde van Vlaanderen Results - K-Means Clustering

This study applies the K-Means clustering algorithm to analyze the results of cyclists in various iterations of the Ronde van Vlaanderen race. The clustering process utilized 11 features: `birth_year`, `weight`, `height`, `climb_total`, `cyclist_age`, `cyclist_bmi`, `climb_percentage`, `race_physical_effort`, `previous_mean_position`, `previous_mean_delta`, and `cyclist_previous_experience`. These features were selected following a comprehensive data understanding process, including data cleaning and feature engineering. Dimensionality reduction techniques such as UMAP and PCA confirmed their alignment with the principal components, ensuring their relevance for clustering analysis.

4.9.1 Cluster Analysis

To determine the optimal number of clusters, the Elbow Method and Silhouette Score were applied. This analysis suggested that a configuration of 4 clusters provided a good balance between compactness and separation, yielding meaningful and interpretable results.

Cluster results were examined using 2D and 3D visualizations to analyze the defining characteristics of each cluster (Figures 22a 22b).

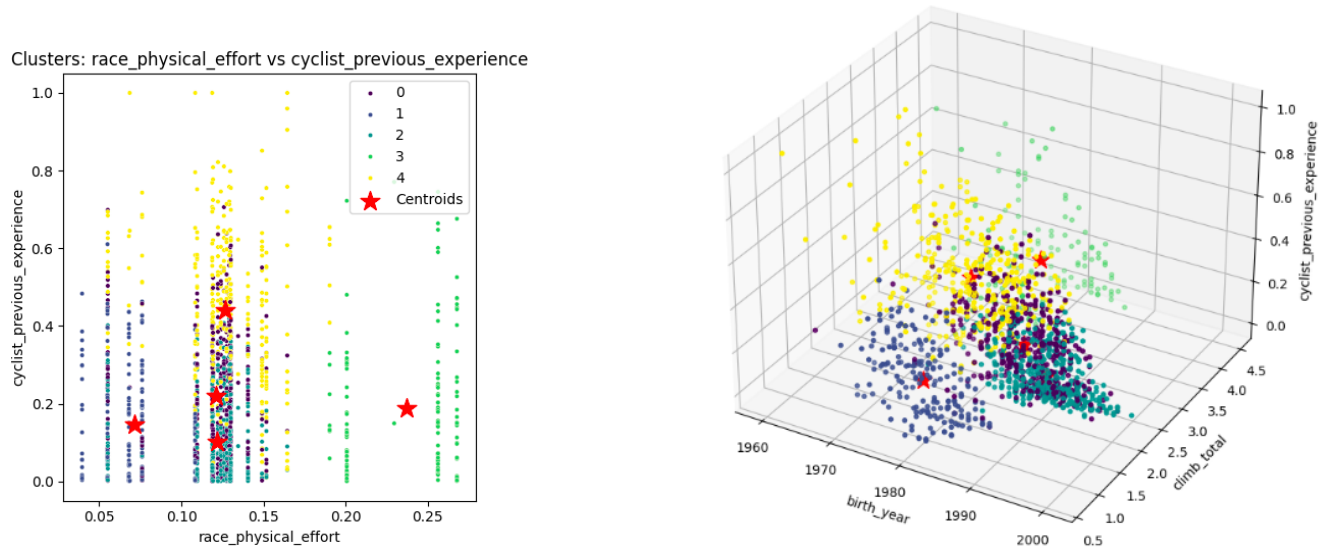


(a) 2D plot

(b) 3D plot

Figure 21: Plots showing the main results of K-Means applied to the Races data.

Despite the inclusion of a broad set of features, the clustering revealed limited differentiation between meaningful subgroups, particularly concerning the combination of race-related and cyclist-specific features.



(a) 2D plot

(b) 3D plot

Figure 22: Plots showing the main results of K-Means applied to the Ronde van Vlaanderen race data.

4.9.2 Observations

Among the 4 identified clusters, most showed relatively homogeneous patterns with minimal feature mixing. The only moderately distinct cluster was **Cluster 4 (yellow)**. This cluster tended to include cyclists with higher values of `cyclist_previous_experience`, though the observed distinction was weak and did not provide clear insights into broader patterns. The overall clustering did not reveal meaningful or interpretable patterns. Future research could explore alternative clustering techniques or include additional domain-specific features to enhance the interpretability of the results.

5 Classification

The classification task focused on predicting whether a cyclist achieved a top-20 placement given characteristics about race and the cyclist. To achieve this, we partitioned the dataset into a development set, comprising races held before 2022, and a test set, which included races from 2022 onward. The target labels were defined as binary: 1 for top-20 placement and 0 otherwise.

5.1 Data Preprocessing

Feature selection was a crucial step in preparing the data for classification. Several features were excluded to enhance the generalizability and relevance of the model. Features such as `cyclist_url`, `cyclist_name`, `race_url`, `race_name`, `dates`, `race_year`, `race_stage`, and `cyclist_team` were removed due to their lack of generalizability. Features like `birth_year` (already captured by `age`), `weight` and `height` (combined into BMI), and `uci_points` (incorporated into points) were omitted due to redundancy. Similarly, features like `points` and `startlist_quality` were considered as consolidated into `race_prestige`, while `climb_total` and `profile` were considered encapsulated in `race_physical_effort`. Features such as `nationality` and `mostly_tarmac` were excluded due to imbalance in their distributions. Additionally, `raw_position`, `position`, `delta`, and `cyclist_climb_power` were removed as they contained label-related information.

When choosing between the features `cyclist_age` and `cyclist_age_group`, we opted to retain `cyclist_age_group` since it was engineered to ensure a better distribution in the balanced distribution of its values in the dataset. The final set of features included: `length`, `race_season`, `cyclist_bmi`, `cyclist_age_group`, `climb_percentage`, `race_physical_effort`, `race_prestige`, `previous_mean_position`, `previous_mean_delta`, `previous_mean_cp` and `cyclist_previous_experience`, and `num_participants`

The categorical attributes were transformed into numerical equivalents to facilitate processing by machine learning models. For non-ordinal categorical attributes, such as `race_season` one-hot encoding was applied. This approach maximized dissimilarity between classes by treating each category as a separate binary feature. For ordinal attributes, such as `cyclist_age_group` discretization was used to preserve the inherent order and proximity relationships among the classes.

5.2 Dataset Balancing

The dataset exhibited a significant imbalance between the two classes, necessitating the application of balancing techniques. Two balanced datasets were created to evaluate model performance under different conditions. The first was an oversampled dataset, constructed using a Random OverSampler. This technique was selected after comparing it with other methods, such as SMOTE, SMOTENC, and ADASYN, as it best replicated the original data distribution. The second was an undersampled dataset, created using a Random UnderSampler. This method was chosen for its computational efficiency, particularly given the long execution times of alternative undersampling techniques.

5.3 Model Selection and Validation

To ensure robust performance, we explored a diverse set of machine learning models and conducted thorough model selection. The evaluation process was tailored to the characteristics of the datasets. For the oversampled dataset, Hold-Out Cross Validation was employed. This approach involved separating the validation and training sets before applying oversampling, thus ensuring that no data were duplicated or shared between the sets. For the undersampled dataset, K-Fold Cross Validation was used, leveraging the data more effectively while preventing overfitting.

Starting from the processed datasets, we trained and evaluated several classes of machine learning models. These included Decision Tree, Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbors (KNN), Random Forest, XGBoost, AdaBoost, Neural Networks, and Rule-Based models.

The Neural Network (NN) architecture was customized using Keras, encapsulated within a hypermodel for hyperparameter tuning with `keras_tuner`. This framework allowed for flexibility in adjusting key parameters to optimize model performance.

The network structure comprised multiple layers:

- **First Layer:** A fully connected Dense layer with a customizable number of units and a ReLU activation function to introduce non-linearity.
- **Dropout Layer:** Incorporated to mitigate overfitting by randomly deactivating a fraction of neurons during training, with the dropout rate adjustable as a hyperparameter.
- **Second Dense Layer:** Another fully connected layer with half the number of units as the first layer (`units//2`), maintaining the ReLU activation function.
- **Output Layer:** A single neuron with a sigmoid activation function, suitable for binary classification tasks.

The model was compiled using the Adam optimizer, with the learning rate treated as a tunable hyperparameter. The loss function chosen was binary cross-entropy, tailored for binary classification.

The model selection process was conducted in two rounds, leveraging the oversampled and undersampled datasets:

1. **First Round:** Randomized Search was used to explore the hyperparameter space and identify approximately three promising models for each dataset.
2. **Second Round:** Bayesian Optimization was employed to refine the hyperparameter tuning for the shortlisted models. This approach, efficient for expensive evaluations, utilized a probabilistic surrogate model to iteratively improve parameter selection. By balancing exploration and exploitation, Bayesian optimization minimized the computational cost of identifying optimal configurations.

Three performance metrics guided the model selection: F1 Score for class 1 (top-20 placements), F1 Score for class 0 (non-top-20 placements), and the Macro Average F1 Score, ensuring balanced performance across classes.

The best-performing models for each dataset were as follows:

Class	Precision	Recall	F1-Score
0	0.94	0.68	0.79
1	0.26	0.73	0.38
Accuracy		0.68	
Macro Avg.	0.60	0.70	0.58
Weighted Avg.	0.85	0.68	0.73

Table 1: Performance metrics for XGBoost on the undersampled dataset, calculated on Test Set after a retraining phase on the whole Development Set

Class	Precision	Recall	F1-Score
0	0.90	0.92	0.91
1	0.42	0.36	0.38
Accuracy		0.85	
Macro Avg.	0.66	0.64	0.65
Weighted Avg.	0.84	0.85	0.84

Table 2: Performance metrics for Random Forest on the oversampled dataset, calculated on Test Set after a retraining phase on the whole Development Set

6 Explainability

Given the poor results of the classification task, the goal is to understand the reasons for this outcome. Specifically, this involves analysing the performance of the winning models from both the undersampling and oversampling approaches to identify patterns, limitations or insights that might explain the observed behaviour.

6.1 Undersampling

To understand the performance of the classifier on the undersampled dataset, SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations) and LORE (LOcal Rule-based Explanations) are used. Additionally, other methods, such as counterfactual explanations and iterative feature removal, were also employed; however, the results from these methods are not included in this report.

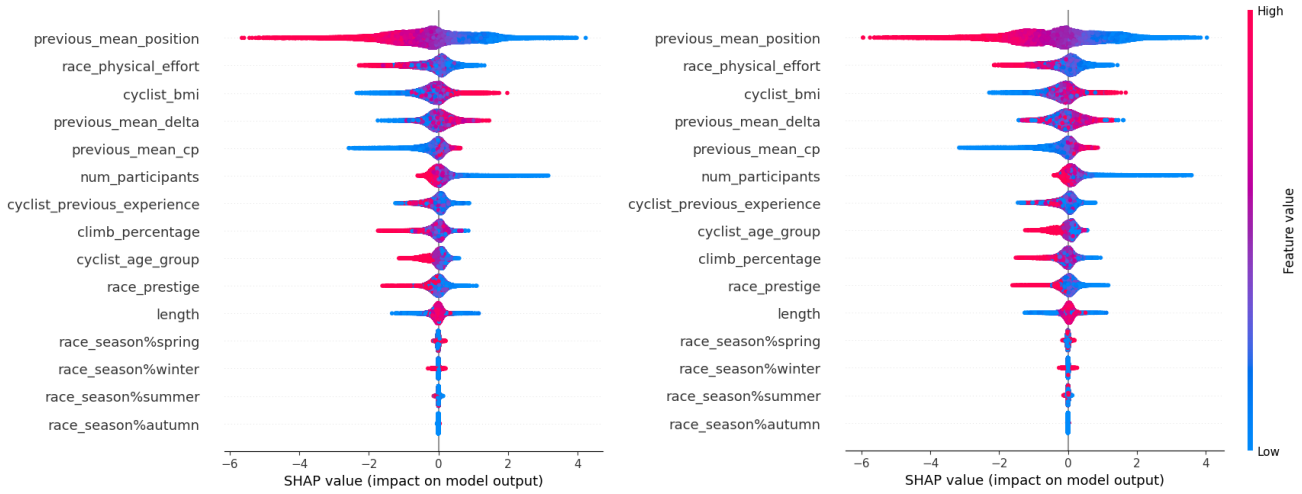
6.1.1 SHAP

The analysis of the interventional and distributional explanations, as shown in Figure 23, shows that the feature `previous_mean_position` is the most important for classification. This is evident from its extreme SHAP values, which indicate its strong influence on the model's decisions. Specifically, low values of `previous_mean_position` push the decision towards class 1 (top 20), while high values result in strongly negative SHAP values, pushing the decisions towards class 0 (not top 20). Other features such as `race_physical_effort`, `num_participants`, `climb_percentage`, `cyclist_age_group` and `race_prestige` show a similar relationship with classes 0 and 1 as `previous_mean_position`. However, these features have smaller SHAP values, making them less influential in the classification process. In contrast, features such as `previous_mean_delta`, `previous_mean_cp` and `cyclist_bmi` show an opposite pattern to `previous_mean_position`. High values of these features are associated with decisions in favour of class 1, while low values influence decisions in favour of class 0. Finally, some features, including `race_season`, show minimal influence on the decision process, as indicated by SHAP values close to zero and no discernible relationship with class 1 or 0 classifications.

6.1.2 LIME

The LIME method is used to perform a local analysis of the model's decisions, focusing on the cases that resulted in true positives, true negatives, false positives and false negatives.

Examining the LIME plots in Figures 24, it is evident that the `previous_mean_position` feature plays a significant role in determining the class of these instances. Only in 24b this feature seems to have a less significant role, maybe due to the fact that its value is 0.54, which is the median of the ranges of this feature (ranging from 0 to 1). The other features appear to have a medium to small influence compared to the `previous_mean_position` feature, a finding that is also supported by the SHAP analysis.



(a) Distributional Explanation.

(b) Interventional Explanation.

Figure 23: SHAP analysis.

6.1.3 LORE

The LORE analysis took a similar approach to LIME, looking at explanations for examples of false positives, false negatives, true positives and true negatives (the same examples as those used in LIME).

The mimic-decision-tree generated by LORE reveals the following crucial decision nodes for each type of classification result: Here's the revised text in English, formatted as a single block without bullet points but retaining the format of the terms and slightly summarizing each explanation:

- For a **true positive instance**, key factors include `previous_mean_position` $\leq$ `0.39`, indicating strong past performance within the top 40% of participants, and `previous_mean_delta` $\leq$ `124.69`, showing an average gap of under two minutes behind the leader, reflecting competitiveness. A physical effort greater than `0.11` (`race_physical_effort` $>$ `0.11`) underscores the athlete's ability to handle intense races. Despite very low prior experience (`cyclist_previous_experience` $\leq$ `0.01`), predictions remain favorable. Additionally, an age factor of `cyclist_age_group` $\leq$ `1.51` may contribute to resilience and recovery, supporting positive outcomes.

- For **true negative instances**, a high `previous_mean_position` $>$ `0.66` suggests weaker past results, reducing the likelihood of success. Similarly, a longer race (`length` $>$ `34.00`) might highlight gaps in endurance or strategy, making it more challenging for the cyclist to compete effectively.

- The **false positive instances** reveal that cyclists with a `previous_mean_position` $\leq$ `0.41` have shown strong prior placements, though `cyclist_previous_experience` $\leq$ `0.51` indicates limited experience that could skew predictions. A race effort within the range `0.02 < race_physical_effort` $\leq$ `0.35` indicates moderate intensity, which might be manageable for the cyclist. The presence of climbing (`climb_percentage` $>$ `0.00`) and course lengths exceeding `32.50 km` suggest the exlusion of very low-level challenges. The average time difference (`previous_mean_delta` $\leq$ `407.84`) shows that the cyclist is relatively close to the top performers, though not definitively exceptional.

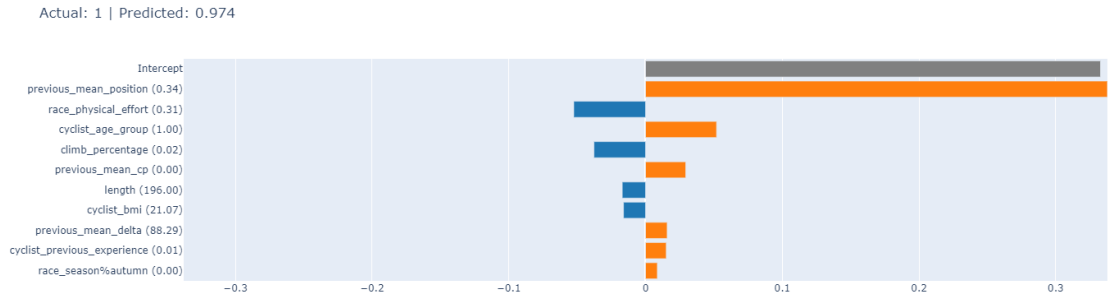
- For **false negatives instances**, a `previous_mean_position` $>$ `0.53` reflects inconsistent placement among the best competitors. Crowded races (`num_participants` $>$ `40.5`) add further difficulty, with success requiring top 50% placement. Although a gap under `598.62 seconds` (10 minutes) indicates closeness to leaders, a `previous_mean_delta` above `154.30 seconds` (2.5 minutes) still points to challenges in staying with the front pack. Additionally, a body mass index (`cyclist_bmi` $>$ `19.01`) above this threshold may be suboptimal for top-level competition.

6.2 Oversampling

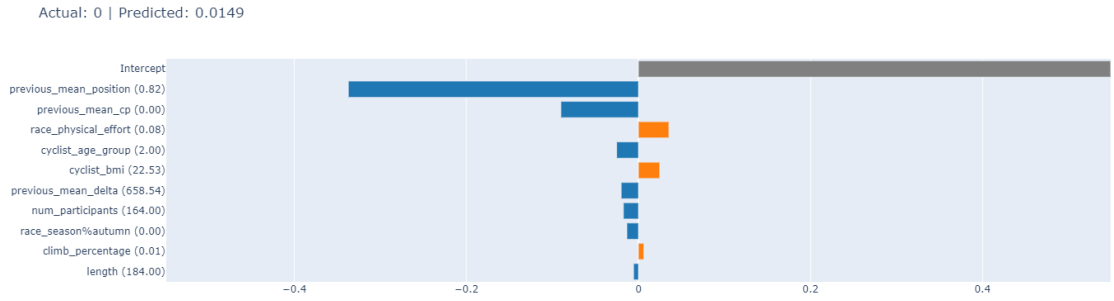
To understand the performance of the classifier on the oversampled dataset, LIME and LORE are used. SHAP was not used due to technical problems encountered during its execution, in particular a segmentation error. Additionally, other methods, such as counterfactual explanations and iterative feature removal, were also employed; however, the results from these methods are not included in this report.

6.2.1 LIME

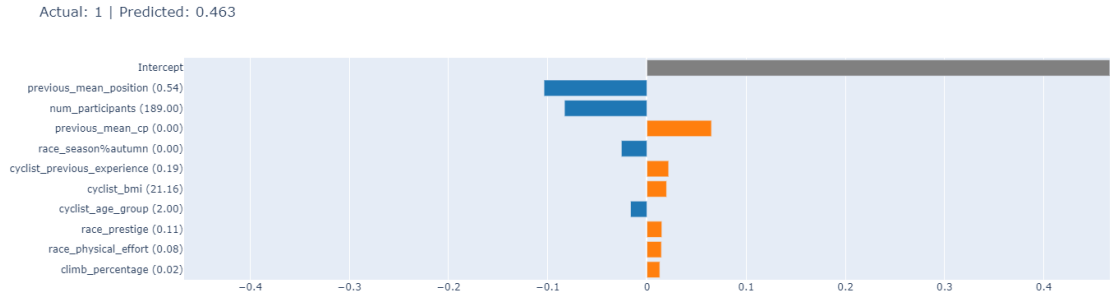
The LIME procedure was applied in the same way as in the previous analyses. Looking at the LIME plots in Figures 25, it is clear that the `previous_mean_position` suggests a similar behaviour to the previous LIME analysis.



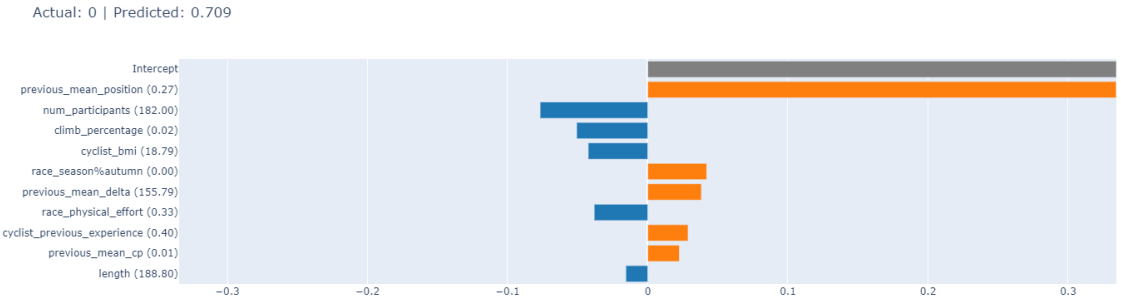
(a) LIME true positive.



(b) LIME true negative.



(c) LIME false negative.



(d) LIME false positive.

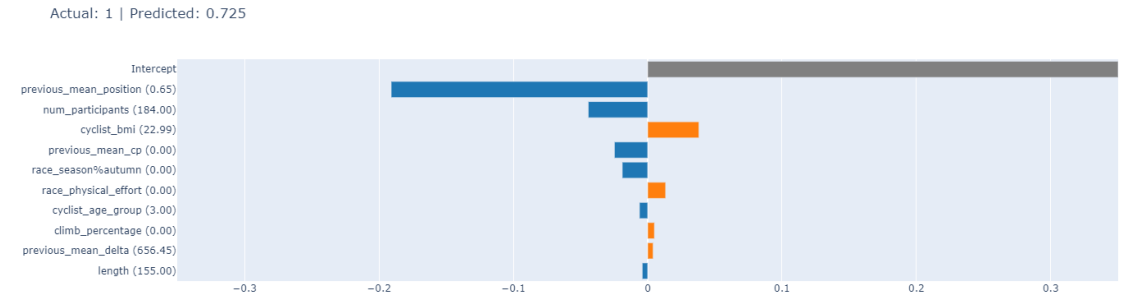
Figure 24: LIME analysis.

6.2.2 LORE

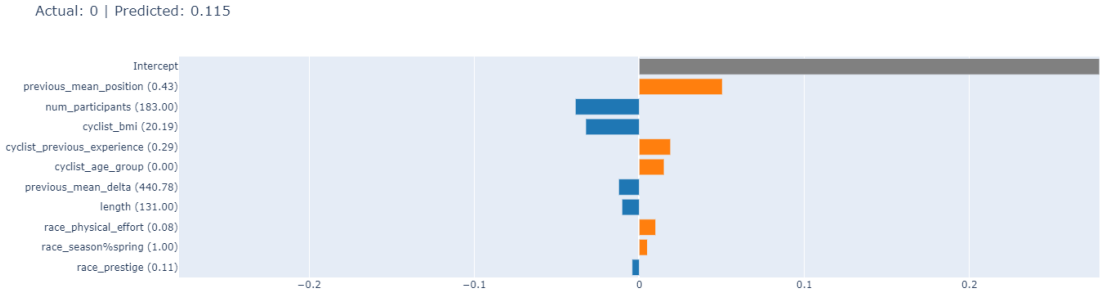
The LORE procedure was applied in the same way as in previous analyses. The mimic-decision-tree generated by LORE reveals the following crucial decision nodes for each type of classification result:

- For **true positive instances**, the cyclist's BMI falls between $21.45 < \texttt{cyclist\_bmi} \leq 22.99$, indicating a lean yet muscular physique suitable for endurance races. Despite a $\texttt{previous\_mean\_position} > 0.37$, ranking in the third quartile, the cyclist performs well in large competitions with over 168 participants ($\texttt{num\_participants} > 168.00$).Minimal elevation ($\texttt{climb\_percentage} \leq 0.01$) indicates flat race conditions. The model incorporates this information, along with the observation that the cyclist has not previously shown climbing capabilities ($\texttt{previous\_mean\_cp} \leq 0.00$), considering it either an advantage or at least not a disadvantage in the classification.

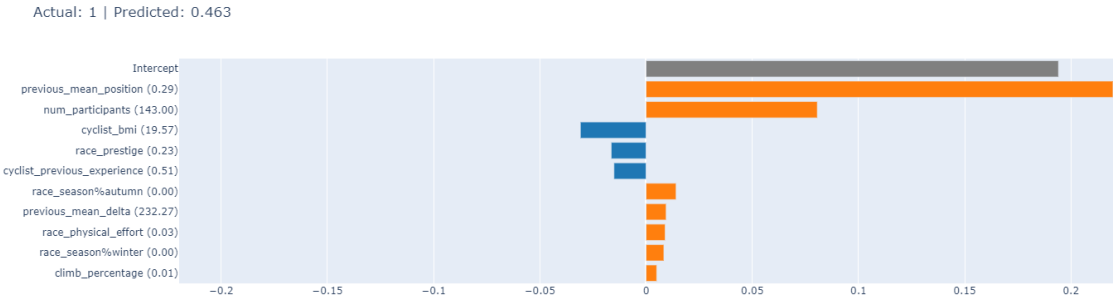
- For **true negative instances**, the cyclist's $\texttt{previous\_mean\_position} > 0.43$ reflects placements in the middle of the ranking, with a BMI of $\texttt{cyclist\_bmi} \leq 21.44$ potentially highlighting limited strength. A moderate number of competitors ($\texttt{num\_participants} > 96.50$) and a significant time gap of over 224.87 seconds from the leader ($\texttt{previous\_mean\_delta} > 224.87$) further suggest performance struggles in competitive contexts.



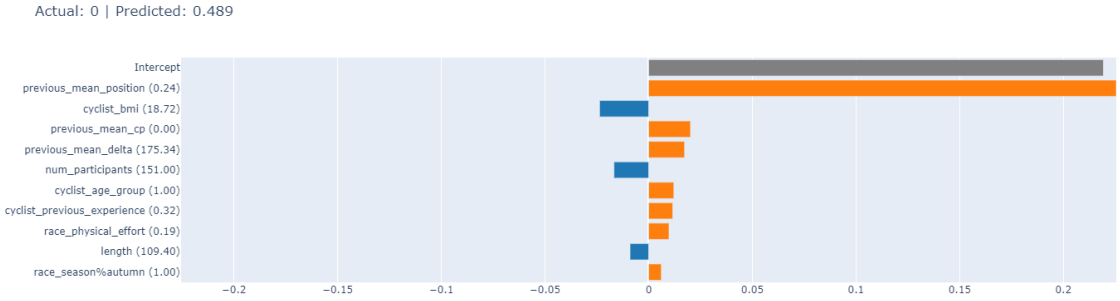
(a) LIME true positive.



(b) LIME true negative.



(c) LIME false negative.



(d) LIME false positive.

Figure 25: LIME analysis.

- For **false positive instances**, the cyclist's `previous_mean_position = 0.55` reflects consistent placement in the top 55%, while a low `race_prestige` $\leq$ `0.27` and minimal `race_physical_effort` $\leq$ `0.07` suggest a low-relevance, low-intensity event. A gap of `previous_mean_delta` $>$ `593.53 seconds` indicates decent competitiveness despite maintaining some distance from leaders. Limited `cyclist_previous_experience` $\leq$ `0.19` highlights minimal past race exposure. However, a favorable `cyclist_bmi` $>$ `20.78` points to optimal physical condition. Finally, `climb_percentage = 0` confirms the cyclist's performance is restricted to flat races.
- For **false negative instances**, a `previous_mean_position` $>$ `0.27` suggests mid-tier rankings, and a large number of competitors (`num_participants` $>$ `124.50`) adds competition. The `cyclist_bmi` $\leq$ `21.38` may provide advantages in climbing but could limit strength for other races. The race's physical effort is very low (`0.00` $<$ `race_physical_effort` $\leq$ `0.06`), which, combined with the cyclist's higher BMI, could pose a disadvantage. Additionally, the race prestige exceeds `0.04` (`race_prestige` $>$ `0.04`), marking it as a moderately significant event, and for a less prepared cyclist, this factor may act as a neutral or even negative influence on performance.

In conclusion, after analysing the performance of the classifiers on both the undersampled and oversampled datasets, it can be stated that the main reason for the poor performance of the classifier may be its heavy reliance on the

`previous_mean_position` feature for decision making. This feature appears to play a dominant role, potentially overshadowing other important factors and leading to suboptimal classification results.

7 Conclusion

This project demonstrates the application of data mining methodologies to analyze and model cyclist and race data, producing valuable insights and identifying areas for further research. Starting with data understanding, the study uncovered patterns such as the strong correlation between cyclist height and weight, and the lack of significant relationships between most race attributes. These findings highlight the importance of thorough initial analyses to guide subsequent steps effectively.

Preprocessing and feature engineering proved critical for ensuring data quality and model relevance. Techniques like feature scaling, imputation, and the introduction of attributes such as `race_prestige` and `cyclist_previous_experience` enabled a more nuanced analysis. The clustering phase revealed meaningful groupings, such as clusters distinguished by performance metrics or race difficulty, providing insights into the characteristics of cyclists and races.

The classification models, although limited in their predictive power, underscored the dominance of specific features like `previous_mean_position`, which played a pivotal role in the models' decisions. Explainability tools further illuminated these patterns, revealing that the reliance on a small subset of features might have hindered the models' generalization capabilities. These findings point to opportunities for refining the feature set, improving dataset balance, and exploring alternative modeling approaches to enhance predictive accuracy.

Overall, this project explores the power and limitations of data mining techniques in analyzing complex, real-world datasets. By combining robust preprocessing, advanced machine learning methods, and interpretability frameworks, the study explored in detail a typical roadmap for research and practical applications in sports analytics.