MASTERS

Quantitative Methods for Economic and Business Decision

MASTER'S FINAL WORK

PROJECT

Pacing Strategies in 800m and 1500m Freestyle: A Data-Driven Analysis from the 2024 Olympic Games

JOANA AFONSO PINTO

JULY - 2025

MASTERS

Quantitative Methods for Economic and Business Decision

MASTER'S FINAL WORK

PROJECT

Pacing Strategies in 800m and 1500m Freestyle: A Data-Driven Analysis from the 2024 Olympic Games

JOANA AFONSO PINTO

SUPERVISOR:
PROF. CARLOS J. COSTA

JULY - 2025

Acknowledgements

For my mother, who, though not here to witness my achievements, continues to inspire and guide me. Every step I take carries a piece of her.

To my father, no words of thanks can ever match all he has given me, his support, his sacrifices and his unconditional love.

My thanks also to Anabela, who never gave up on me and her support played a profound role in shaping the person I have become, even if she may not fully realise it.

To my family and friends, who have been my pillar through every challenge.

Finally, to Diogo, who never let me feel alone, and spent countless hours by my side, listening and supporting me every step of the way.

Resumo

Compreender o ritmo dos nadadores de elite durante as competições é essencial para melhorar o seu desempenho em provas de longa distância. Este estudo explora as estratégias de ritmo adotadas pelos atletas nas provas de 800m e 1500m livres nos Jogos Olímpicos de 2024, com um foco específico na identificação dos fatores que explicam as variações da velocidade de natação e na avaliação de qual o modelo de machine learning que melhor as prevê.

Inicialmente, foi considerada uma abordagem baseada na classificação, com o objetivo de prever perfis de ritmo a partir de características da prova. No entanto, devido ao tamanho e à natureza limitada do conjunto de dados, esta abordagem foi descartada. Como alternativa, adotou-se uma metodologia em duas etapas: (i) foram exploradas estratégias de ritmo através de *agrupamento hierárquico aglomerativo*; e (ii) foram utilizados modelos baseados em regressão para explicar e prever a velocidade do nadador ao longo da prova.

A análise de agrupamentos revelou três perfis distintos de ritmo nos 800m, duas estratégias em forma de U (um mais rápido e outro mais lento) e uma estratégia de ritmo positivo, enquanto dois perfis em forma de U foram identificados nos 1500m. Os testes estatísticos confirmaram que estes agrupamentos estavam associados ao sexo, ao tempo de entrada e à variabilidade do ritmo (CV%), mas não à classificação final da prova.

Para estudar os determinantes da velocidade, foram geradas novas variáveis, incluindo a aceleração, a distância até à chegada e o tempo do parcial anterior. A análise de importância das características identificou o sexo, a aceleração e o tempo de entrada como os preditores mais fortes. Entre os modelos testados, o *Gradient Boosting* apresentou o melhor desempenho preditivo, superando o *Random Forests*, as *Redes Neurais* e a *regressão OLS*. A análise dos resíduos, incluindo o teste de *Durbin-Watson*, confirmou a robustez estatística dos modelos.

Palavras-Chave: estratégia de ritmo, natação, previsão de velocidade, machine learning, Jogos Olímpicos, modelos de regressão.

Abstract

Understanding how elite swimmers pace themselves during competition is essential for improving performance in long-distance events. This study explores the pacing strategies adopted by athletes in the 800m and 1500m freestyle races at the 2024 Olympic Games, with a particular focus on identifying the factors that explain variations in swimming velocity and evaluating which machine learning model best predicts it.

Initially, a classification-based approach was considered, aiming to predict pacing profiles from race features. However, due to the limited size and nature of the dataset, this approach was discarded. As an alternative, a two-step methodology was adopted: (i) pacing strategies were explored through agglomerative hierarchical clustering; and (ii) regression-based models were used to explain and predict swimmer velocity throughout the race.

The clustering analysis revealed three distinct pacing profiles in the 800m, two U-shaped patterns (one faster and one slower) and one positive-split strategy, while two U-shaped profiles were identified in the 1500m. Statistical tests confirmed that these clusters were associated with sex, entry time, and pacing variability (*CV%*), but not with final race ranking.

To study the determinants of velocity, new variables, including *acceleration*, *distance to the finish line,* and *previous split*, were computed. Feature importance analysis identified sex, acceleration, and entry time as the strongest predictors. Among the models tested, *Gradient Boosting* revealed the best predictive performance, outperforming *Random Forests*, *Neural Networks*, and traditional *OLS regression*. Residual analysis, including the *Durbin-Watson test*, confirmed the statistical robustness of the models.

Keywords: pacing strategy, swimming, velocity prediction, machine learning, Olympic Games, regression models.

Index

Table Index

Figure Index

Annex Index

Glossary

AI – Artificial Intelligence

ANN – Artificial Neural Network

CV (%) – Coefficient of Variation

DT – Decision Tree

FAMD – Factor Analysis of Mixed Data

GB – Gradient Boosting

ML – Machine Learning

MLP – Multi-Layer Perceptron

MAE – Mean Absolute Error

MSE – Mean Squared Error

OLS – Ordinary Least Squares

PR – Precision-Recall

RF – Random Forest

RMSE – Root Mean Square Error

RFE – Recursive Feature Elimination

SSB- Sum of Squares Between groups

SSE – Sum of Squares Error

SVM – Support Vector Machine

1 Introduction

Pacing, the distribution of effort across a race, is a critical determinant of performance in endurance sports, especially swimming. In long-distance events like the 800m and 1500m freestyle, optimal pacing can differentiate between reaching the podium or not. Recent literature identifies the U-shaped pacing strategy as a dominant pattern among elite swimmers, often characterised by fast starts and finishes, with a relative slowdown in the middle of the race.

Although pacing has been extensively studied in sports like running and cycling, swimming remains comparatively underexplored, particularly from a data science perspective. Traditional performance analysis in swimming has relied on basic statistics and descriptive metrics. However, advances technologies have increased the availability of data. In this specific case, granular split-time data opens new opportunities to leverage machine learning techniques to gain a deeper understanding of the athlete's behaviour and race dynamics.

The main objectives of this study are to understand the pacing profiles adopted by elite swimmers in the 800m and 1500m freestyle events at the 2024 Olympic Games, to assess the key factors that influence their in-race velocity and to identify the most effective machine learning model for predicting the swimmers' velocity. Initially, the goal was to predict each athlete's pacing strategy through classification models using performance-related features. However, due to the small dataset, this classification approach was ultimately discarded.

This study begins with a literature review aimed at exploring previous research on pacing strategies in swimming, the application of machine learning techniques in sports performance analysis, and an in-depth review of the techniques that were applied throughout the study. Next, the data and methodology are presented, including a description of the variables used and the preprocessing steps. The methodological process can be divided into three main phases: (i) the first phase involves exploratory and unsupervised analysis to identify and characterize the pacing profiles adopted by athletes; (ii) the second phase focuses on determining the key features that influence swimming velocity, supported by regression modelling; (iii) in the third phase, various predictive models are tested and evaluated to determine which algorithm best explains swimming velocity, considering model accuracy and robustness. Following this, the discussion of the results obtained in the different stages is discussed and interpreted in light of the revised literature. Finally, the conclusion summarises the main findings of the study and proposes directions for future research.

2 Literature Review

This chapter includes a thorough review of the existing literature on pacing strategies in long-distance freestyle swimming, particularly the 800 and 1500-meter events. It further explores the application of statistical methods and highlights the growing importance of machine learning in sports, pointing out its limited application in swimming. The aim is to uncover the pacing strategies used by elite swimmers in their races and to determine the key factors that influence these strategies.

2.1 Machine Learning

Artificial intelligence (*AI*) can be defined as a set of systems that enable machines to have human-like intelligence, including the ability to learn, perceive, reason, and interact (Russell & Norvig, 2022).

Machine learning (*ML*), a core subset of *AI*, is the analysis of algorithms that allows systems to learn and enhance their performance based on experience (Sah, 2020). *ML* has three primary approaches: supervised, unsupervised, and reinforcement learning (Figure 1). The key difference between supervised and unsupervised learning is the type of data used. Supervised learning uses labelled datasets, allowing the algorithms to learn to classify data points or predict outcomes with improved accuracy over time (Kotsiantis, 2007). This approach includes two major categories: regression (for continuous outputs) and classification (for discrete outputs) (Bousquet et al., 2004). On the contrary, unsupervised learning models work with unlabelled data to uncover hidden patterns without any human intervention (Mahesh, 2020). Typical tasks for this approach include clustering, association, and dimensionality reduction.
Reinforcement learning involves an agent that learns to make decisions by interacting with an environment. The agent obtains feedback in the form of rewards or penalties for the actions it performs, being the goal of maximising the total reward (Sah, 2020).

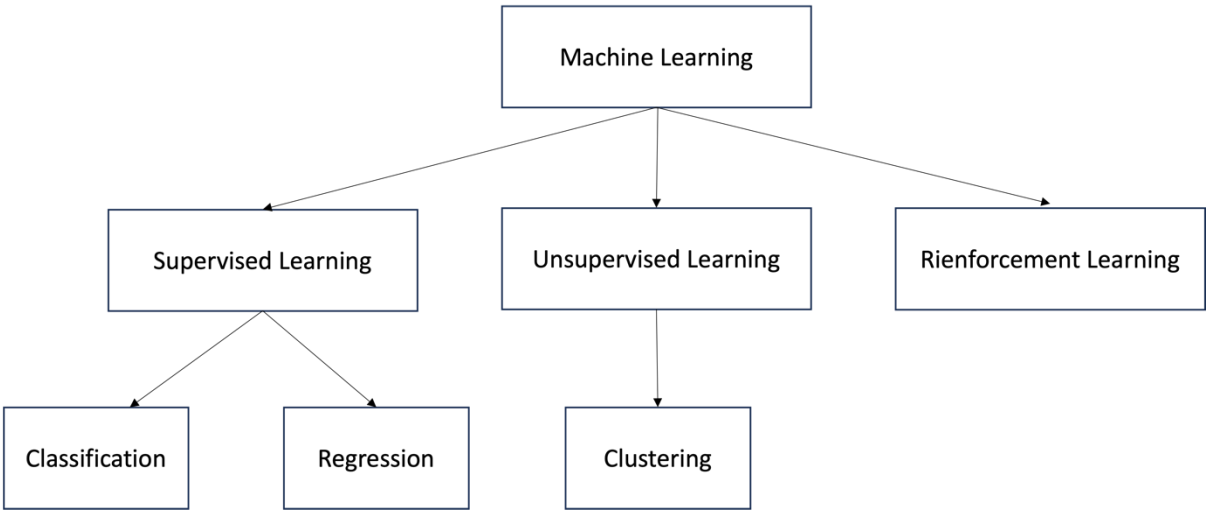


Figure 1: Types of Machine Learning

Given the variety of approaches in *ML*, selecting the appropriate models depends on the characteristics of the problem and the structure of the data. This study used both supervised and unsupervised learning techniques to explore different aspects of swimmer performance.

Unsupervised learning was used to identify distinct pacing strategy profiles, while supervised models were applied not only to attempt classification of those strategies but also to explain and predict velocity during the race, a continuous variable closely related to pacing behaviour. To address these objectives, a combination of advanced algorithms, including Neural Networks, Decision Trees, Support Vector Machines, and ensemble models[1] such as Random Forest and Gradient Boosting, were implemented. The following sections describe each of these methods.

2.1.1 Unsupervised Algorithms

As mentioned, *unsupervised learning* is a branch of *ML* used to detect hidden patterns and groupings in datasets that do not have predefined labels.

When the dataset to be analysed contains both categorical and continuous variables, traditional *Principal Component Analysis (PCA)* is not suitable for dimensionality reduction. In such cases, ***Factorial Analysis of Mixed Data (FAMD)*** can be an effective alternative (Audigier et al., 2016). *FAMD* is a key component technique for summarising and characterising mixed data, primarily intended to investigate individual similarities, the connections among variables, and to relate the analysis of individuals to that of the variables. It can be viewed approximately as a combination of *PCA* and *Multi Correspondance Analy*sis (*MCA)*. Specifically, the continuous variables are scaled to unit variance while the categorical variables are converted into a disjunctive data table and afterwards scaled according to the specific scaling of *MCA* (Pagès, 2014).

One of the most common tasks is ***Clustering***, which seeks to classify different data points based on their similarities or patterns. A *Hierarchical Clustering* method forms groups (clusters) by recursively dividing the instances in either a top-down or bottom-up manner. This technique can be divided into *Agglomerative* and *Divisive* hierarchical clustering (Rokach & Maimon, 2005).

For this study, the *Agglomerative* technique was used. In this technique, each object first represents a cluster of its own, then clusters are gradually combined until the desired cluster structure is achieved. To measure similarity between instances, *Gower distance* is the most suitable for datasets containing mixed data types, as it combines different distance metrics depending on the nature of each attribute. The overall dissimilarity is computed as an average of the individual attribute distances (Liu et al., 2024):

$$d(x_i, x_k) = \frac{1}{p}\sum_{j=1}^{p} d_j\left(x_{ji}, x_{jk}\right)$$

where p is the number of variables, and $d_j\left(x_{ji}, x_{jk}\right)$ is the distance between 2 observations, which is computed different whether the variables is continuous or categorical (Liu et al., 2024).

To compute the distance between two clusters, there are several options, including *Single Linkage, Average Linkage, Complete Linkage* and *Ward's method* (Miyamoto, 2022). *Average*

[1] Ensemble leaning refers to the technique of combining multiple models to produce a more robust predictive outcome (Mahesh, 2020).

Linkage defines the distance between two clusters as the average of all pairwise distances between the elements of each cluster. This method is particularly appropriate in conjunction with the *Gower distance*.

Another method commonly used for continuous data is *Ward's method,* which minimises the total within-cluster variance, and it considers the loss of information that occurs when clustered together. The key measurement used is the *Error Sum of Squares (ESS)*, which calculates the squared differences between each instance and the centroid (mean) of its cluster. It can be represented as:

$$ESS = \sum_{i=1}^{nclust} \sum_{j=1}^{n} \sum_{k=1}^{v} \left(X_{ijk} - \bar{X}_{i \cdot k}\right)^2,$$

where:

- X_{ijk} is the value of the variable k for the instance j in the cluster i.
- $\bar{X}_{i \cdot k}$ is the mean value of the variable k within a cluster i.
- v is the number of variables

2.1.2 Supervised Algorithms

Supervised algorithms are *ML* models that learn the relationship between input features and known outputs. Depending on the nature of the task and structure of the data, these models can be used for either classification or regression.

One of the simplest to understand supervised machine learning algorithms is ***Ordinary Least Squares (OLS)***, commonly known as ***Linear Regression.*** This model is called simple linear regression when only one independent variable is used. However, when two or more independent variables are involved, it is referred as multiple linear regression (Lindholm et al., 2019). The model can be represented by:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,$$

where Y represents the dependent variable, x_i the independent variables, the coefficients β_i are the parameters of the models and ε the error associated with the observed values for Y.

Decision Trees (DT) is an algorithm that classifies instances by grouping them based on feature values. In a decision tree, each node represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified at the root node and arranged according to their feature value (Kotsiantis, 2007) , as illustrated in Figure 2.

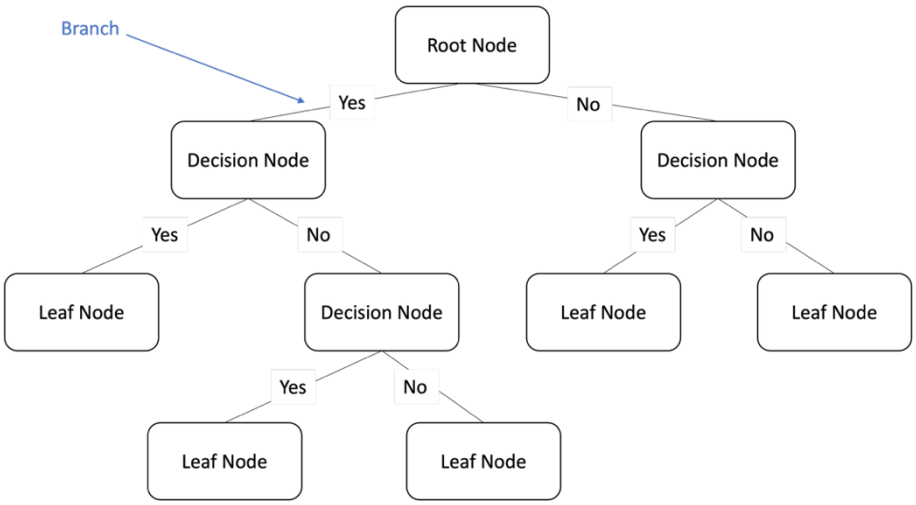


Figure 2: Decision Tree Scheme

Support Vector Machines (SVM) is viewed as a complex algorithm that can provide high accuracy, even when the data sizes are limited (Singh et al., 2016). This algorithm is commonly used for classification problems and can handle both linear and nonlinear classification tasks. *SVM* discriminates between two classes by identifying the best hyperplane that maximises the distance between the nearest data points of opposite classes. The optimal hyperplane can be calculated in the following way:

$$w_0 \cdot x + b_0 = 0,$$

where w is the weight vector, x is the input, and b is the bias term. As various hyperplanes can be classes, maximising the margin between points allows the algorithm to identify the optimal decision boundary between classes. This allows *SVM* to effectively generalise to new data and produce accurate classification predictions.

Radom Forest (RF) is a supervised machine learning algorithm, more specifically, an ensemble learning algorithm, that uses *DT* as its base. This method introduces randomness when building each tree, aiming to create an uncorrelated forest of decision trees, to the bootstrap aggregation method. The bootstrap method selects a random sample from the training data with replacement, and after multiple data samples are generated and the models are trained separately, their predictions are aggregated to produce a final output (Lindholm et al., 2019). The main difference between *DT* and *RF* is that *DT* considers all possible future outcomes, and *RF* only selects a subset of those features. Some of the advantages that *RF* include its robustness to noise, scalability and lower risk of overfitting (Singh et al., 2016). Since the swimming data can have a degree of noise, such as unusually slow times, *RF* can be a great solution to handle these irregularities while still providing useful insights about the data.

Adaptive Boosting (AdaBoost) is also an ensemble learning algorithm that can be used for both classification and regression tasks, although is most commonly applied to classification tasks. This technique is often implemented using decision tree learners and works by consecutively building new models that focus on correcting the errors made by the previous ones. In each iteration, a greater weight is given to the misclassified instance, allowing the model to progressively improve its overall prediction accuracy (Zounemat-Kermani et al., 2021).

Similar to *AdaBoost*, ***Gradient Boosting (GB)*** is an ensemble learning algorithm that combines weak learners to form stronger learners to form a predictive model. Unlike *AdaBoost*, *GB* minimises the loss function by fitting new models to the residuals of the previous ones in an iterative manner. This method can be used for both classification and regression. In each iteration, the algorithm discards weaker predictors and selects the most efficient learners (Bentéjac et al., 2021). The *GB* model can be mathematically expressed as:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x),$$

where F_{m-1} represents the previous model, ρ_m is the weight applied to the m^{th} function, and h_m is the base learner (Bentéjac et al., 2021). To minimise the prediction error, ρ_m is represented by:

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x) + \rho_m h_m(x)\right).$$

The ***Multilayer Perception (MLP)*** is a supervised machine learning algorithm that uses artificial neural networks. It is inspired by the structure of the human brain; it is composed of interconnected layers of nodes, also called neurons (Albon, 2018). *MLP* consists of three main components: an input layer that receives the input features, one or more hidden layers where the data is processed through weight connections and activation functions, and an output layer that generates predictions based on the outputs of the hidden layers (Kotsiantis, 2007), as shown in Figure 3.

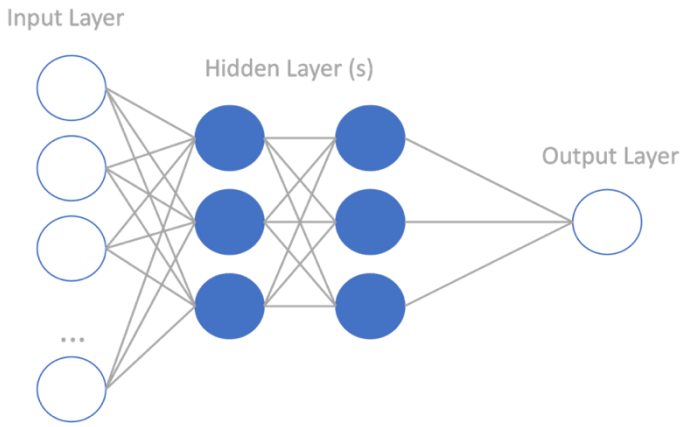


Figure 3: MLP algorithm architecture

2.2 Machine Learning in Sports

In recent years, *AI* has seen a rise in development and adaptation across multiple fields. Once considered a niche topic, tools like ChatGPT are now commonly used daily by millions, demonstrating the integration of AI into our daily lives. This led to a surge of interest among researchers, who now have more to investigate (Collins et al., 2021). One of the areas where researchers have focused with the "bomb" of *AI* is in sports, using machine learning models from making score predictions to predicting athletes' injury risks.

As society has become more performance-driven, the pursuit of excellence in athletics has grown stronger, supported by advances in science and technology. This is where *ML* models can be a big ally.

One of the most notable applications of *ML* in sports is outcome prediction. Various algorithms, such as neural networks, support vector machines (*SVM*), *random forests*, l*ogistic regression*, and *k-nearest neighbours* (*k-NN*), have been applied to predict match results and player performance metrics across sports like football, basketball, and cricket (Horvat & Job, 2020).

In football, a study was conducted to determine whether ball possession affects the goal-scoring likelihood with the objective of assisting coaches in real-time strategy formulation (Markopoulou et al., 2024).

ML is used more and more to identify talent and tailor training. By evaluating physical and psychological data, *ML* can be a valuable tool for coaches to identify young athletes with a high potential of becoming elite athletes, and to customise training programs based on their specific performance profiles (Jauhiainen et al., 2019). For instance, wearable technology integrated with *ML* accuracy and field knowledge allows coaches and sports scientists to respond immediately to biomechanical or physiological changes, enhancing safety and performance (Vec et al., 2024) (Alaguraja & Selvakumar, 2023).
Another critical area is injury risk assessment. *ML* algorithms have been very useful in accessing patterns associated with injuries, using pre-season measures, such as past injuries, training load, and anthropometry measures. These predictive models can help coaches to focus resources in injury risk management as well as give practitioners insights to the specific types of injuries an athlete is more likely to sustain before the start of the season (Rommers et al., 2020).

2.3 Pacing strategies in swimming and their impact on the result

Swimming has seen a notable increase in global interest, both in terms of athlete participation and spectator engagement. The pressure to achieve personal best times and maintain competitive advantages has driven athletes and coaches to find innovative strategies for performance enhancement.

Olympic swimming events involve different phases to reduce the number of competitors. Events are initially divided into sub-events, called heats, and the top swimmers from the heats (usually the top 16) go to the semifinals, where they are split into two heats. The top 8 from the semifinals advance to the final, where the medallists are chosen. For longer events, from 200m up, the 8 fastest times from the heats go through the finals. For each race, the athlete is attributed a lane based on their time from the previous round, with the fastest and second-fastest athletes occupying the middle lanes.

The result of a race can be influenced by many factors, such as stroke efficiency, start and turn performance, pacing strategy, underwater phase, race tactics, lane assignment, physiological factors, competitor awareness, and environmental conditions.

Pacing strategy is a key factor for the result of the race, especially in single sports, and this factor has been amply studied in some sports like cycling, running, but in swimming, there are still few researches that can guide the coaches and athletes on what impact does have pacing strategy and how can they adjust the training to improve this indicator.

Pacing is the rate at which the athlete completes a certain distance. In swimming, this tends to vary by type of event and stroke. Several pacing strategies are commonly observed:

- Negative Split: The swimmer increases speed in the latter half of the race.
- Positive Split: The swimmer starts at a high speed but gradually declines his pace.
- Even Pacing: The swimmer maintains a consistent pace throughout the race.
- All-Out Strategy: Maximum effort from the start, typically seen in sprint events.
- Parabolic or U-Shaped Pacing: Speed decreases in the middle and increases toward the end.
- Variable Pacing: Characterized by inconsistent speed with no discernible pattern; the least used by elite swimmers.

Long-distance events such as 800m and 1500m freestyle require a finely tuned balance between endurance and strategic energy expenditure. As such, the pacing strategy employed by an athlete can significantly influence their final standing (McGibbon et al., 2018). While pacing has been studied across various endurance sports, its influence becomes more pronounced as the event duration increases.

For elite athletes competing in 800m and 1500m freestyle races, the parabolic/U-shape pacing strategy is the most frequently observed. This approach allows for an effective distribution of energy, with reduced speed during the middle phase and a final increase in speed toward the finish. It supports an optimal balance of aerobic and anaerobic energy, conserving anaerobic reserves for a strong final acceleration. In contrast, the swimmers who choose positive pacing (starting fast and gradually slowing) experience higher anaerobic energy depletion early in the race, leading to fatigue and reduced performance in the latter stages (Foster et al., 2003).

While many studies have examined pacing strategies in elite swimmers, few have focused specifically on Olympic data. Given that athletes typically peak at the Olympic Games after four years of targeted training, analysing data from the 2024 Games offers a unique opportunity to assess how pacing strategies correlate with top-tier performance.

Medal-winning swimmers often display less variability in mid-race segments and maintain higher speeds in the final 500 meters, especially in the 1500m event. Stroke frequency and stroke length also emerge as critical indicators to predict pacing stability (Morais et al., 2023).

The position that the athlete takes early in the race is also a determining factor in the likelihood of winning a medal. Being among the top three swimmers by the 600m mark significantly increases the likelihood of winning (Lara & Del Coso, 2021).

Gender also plays a role in pacing dynamics. Men tend to exhibit greater pacing variability across strokes, suggesting reliance on explosive speed at certain points. In contrast, women generally maintain a more even pace throughout the race, reflecting endurance-oriented strategies (Moser et al., 2021).

The coefficient of variation (CV) is commonly used to assess lap-to-lap variability throughout the race. Lower CV values indicate a more stable pacing and are often associated with higher performance levels. In a study conducted for the top 60 all-time ranked 1500m male freestyle swimmers, researchers found that maintaining the speed during the middle phase (500-1000m) is the most critical determinant of success, as swimmers who slow down significantly mid-race struggle to regain speed in the final 500m. The most successful swimmers demonstrated less

lap-to-lap variability (CV), with a balance approach between speed and energy conservation (Hołub et al., 2023).

2.4 From Basic Statistics to Machine Learning in Swimming

Pacing strategy is a crucial factor influencing performance in long-distance freestyle swimming competitions such as 800 and 1500m races. The methodological approaches employed to study pacing behaviours have relied on traditional statistical methods, but there is an increasing interest in applying machine learning techniques to more effectively understand complex, nonlinear connections in race performance data.

The predominant analytical approach found in the literature relies on traditional statistical tests, particularly the use of *Analysis of Variance* (*ANOVA*) alongside Bonferroni post-hoc tests. These techniques have played a crucial role in analysing lap-to-lap pacing variability and in identifying performance differences across different groups, such as elite vs. junior swimmers, male vs. female, and medallists vs. non-medallists (Hołub et al., 2023; Lara & Del Coso, 2021). *ANOVA* allows researchers to determine whether pacing varies significantly among groups and race segments, while post-hoc tests clarify the locations of these differences. For instance, Morais et al. (2023) demonstrated that medallists often exhibit lower lap-to-lap variability and a more significant final push, reinforcing the efficacy of a parabolic pacing strategy.

Furthermore, time-series analysis examined how pacing evolves throughout a race. This approach enabled researchers to classify pacing strategies such as parabolic (U-shaped), positive, negative, and even pacing, depending on the variations in race speed across different segments (Oliveira et al., 2019). Although these methods provide understanding of the time-based allocation of effort, they are fundamentally descriptive and frequently constrained to generalisations at the group level, rather than insights tailored to individual athletes, which is critical for personalised training and feedback.

To address these restrictions, a few recent studies have explored the application of machine learning models, particularly decision trees, to analyse how split-time variables and mid-race speed patterns predict performance outcomes. Oliviera et al. (2019) specifically used *CHAID* (*Chi-squared Automatic Interaction Detection*) algorithm to evaluate pacing data from Olympic finalists. This approach facilitated the identification of key predictive variables, such as mid-race velocity, and their hierarchical role in performance categorisation. Medallists demonstrated considerably more consistent pacing during the race's middle segments, a trend that *CHAID* effectively recognised without relying on linear assumptions.

Nonetheless, machine learning applications in pacing analysis remain rare. The swimming science community continues to rely on traditional statistical models, likely because of their interpretability and historical significance. However, with advancements in data accessibility and computational resources, machine learning methods present a way to achieve more personalised and predictive modelling of racing strategies. They can manage a wider array of variables, identify non-linear relationships, and provide improved predictive accuracy for immediate decision-making in coaching.

In conclusion, although traditional statistical methods have established strong bases for comprehending pacing strategies in swimming, the integration of machine learning can offer a significant opportunity for the future. Most existing studies continue to rely on traditional methods.

3 Methodology

There are many methodologies that researchers use in machine learning projects, but in this case, the ***CRISP-DM*** methodology (*Cross-Industry Standard Process for Data Mining*) will be followed, which is the most common approach. To uncover the pacing strategies that elite swimmers used in the 20204 Olympic Games in the 800m and 1500m freestyle, determine the key factors that influence in-race velocity, and identify which ML model best explains velocity. This process model for data mining consists of six iterative phases: *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* and *Deployment* (Costa & Aparicio, 2020; Schröer et al., 2021).

The first phase is ***Business Understanding***, where the problem and the objective are assessed. The aim of data mining is a crucial phase in this stage, which in this study is to uncover underlying patterns in the data and identify which variables most significantly influence pacing behaviour and performance outcomes.

The next step is ***Data Understanding***, which involves collecting data from various resources and identifying the most suitable data to achieve the study's objective. This includes the data collection, exploring and describing the data, and checking the data quality.

For this study, the data used is from the World Aquatics online site, the information is referent to the 2024 Olympic Games for the swimming races of 800m and 1500m. The 800m event consists of 16 laps of 50 meters, and the 1500m race consists of 30 laps of 50 meters. The original dataset was composed of 13 variables (Table 1).

Table 1: Original variables of the dataset

Name of Variable	Description	Data Type
Event	Race Distance (800m or 1500m)	Continuous
Type	Type of race (e.g, Heat, Final)	Nominal
Lane	Lane assignment in the race	Discrete
Name	Swimmer’s full name	Nominal
Sex	Sex of the swimmer (M/F)	Nominal
Country	The country the swimmer is representing	Nominal
Birthday	Birth date of the swimmer	Discrete
Entry Time	Swimmer’s official entry time before race (seconds)	Continuous
Distance	Distance mark (e.g., 50m, 100m, etc.)	Discrete
Time	Cumulative race time at distance (seconds)	Continuous
Rank	Position at each split distance	Ordinal
Split	Partial Split time for each 50m segment (seconds)	Continuous
Final Rank	Final rank in the event	Ordinal

The dataset contains a total of 2,687 records, categorised by event type, including 384 records from 800m female events, 684 records from 800m male events, 719 records from 1500m female events, and 960 records from 1500m male events. Specifically, within the 800m event, there are 1,008 records, corresponding to 18 female athletes and 31 male athletes, with an average athlete age of 23 years. The mean reaction time is 0.717 seconds, while the average split time per 50 meters is 30.281 seconds. Athletes' ages range from 17 to 39 years old. Reaction times vary significantly, with the fastest recorded reaction of 0.64 seconds and the slowest of 0.81 seconds. Additionally, the fastest split time documented was 25.42 seconds, while the slowest split time was 36.62 seconds.

In the context of the 1500m events, the dataset contains 1,679 records, representing18 female athletes and 24 male athletes. The mean reaction time in these events is approximately 0.72 seconds, ranging from 0.61 seconds to 0.84 seconds. The average split time per 50 meters is 30.84 seconds, with individual times ranging from 26.12 seconds to 34.98 seconds. The youngest athlete is 16 years old, while the oldest is 39, with an overall average age of 24 years.

In the next phase, ***Data preparation***, the selection of the data to be used is conducted, along with the application of various models aimed at enhancing data quality. For this study, the heats and final event data were used. Additionally, new variables were calculated not only to enable a more detailed investigation of pacing strategies but also to capture the dynamics of in-race velocity, therefore providing a foundation for analysing the factors influencing velocity (Table 2).

Table 2: New variables calculated (Feature Engineering)

Name of Variable	Description	Data Type
Age	Age of swimmer on the date when the Olympic swimming events started (27/07/2024).	Discrete
Velocity	Instantaneous velocity (m/s). It corresponds to the difference in splits per race and per type.	Continuous
Acceleration	Change of velocity between splits or over time	Continuous
Coefficient of Variation (CV%)	$CV = \frac{c}{\overline{x}_{velocity}} \times 100,$ where $s_{velocity}$ is the standard deviation and $\overline{x}_{velocity}$ is the mean of velocity. It measures the stability of pacing through the race and allows for assessing lap to lap variability.	Continuous
Speed Variability	Standard Deviation of Velocities.	Continuous
Start Speed	Average velocity during the first third of the race.	Continuous
Middle Speed	Average velocity during the middle third of the race.	Continuous
End Speed	Average velocity during the final third of the race.	Continuous
Final Speed	Velocity during the last 50m of the race.	Continuous
Prev Velocity	Velocity of the previous split for the same swimmer, race and type.	Continuous
Prev Split	Split time of the previous segment.	Continuous
Meters to Finish	Distance remaining to complete the race at each split.	Continuous

Following feature engineering, ***feature selection*** techniques were applied to reduce the number of input variables, improve model interpretability, and prevent overfitting. This technique is the selection of subsets of variables that together have good predictive power. This can be done manually or automatically (Guyon & Elisseeff, 2003). This technique has numerous advantages, including a reduction in training duration, simplification of the models for enhanced interpretability, mitigation of dimensionality and a decrease in the likelihood of overfitting.

The importance of feature selection becomes more evident in instances where irrelevant predictors may negatively impact the model's performance. This is especially true for algorithms such as *SVM* and *Neural Networks*, as well as for *Linear and Logistic Regression models*, which are highly sensitive to correlated predictors. By using these techniques, one can effectively eliminate redundancy and reduce multicollinearity, leading to a more reliable and accurate model.

Feature selection can be divided into three primary methods: wrappers, filters and embedded methods. Each approach offers unique advantages and is suited to different contexts within the modelling process:

- Filter methods are techniques that select features without the use of any machine learning algorithms, select variables as a pre-processing step, independently of the chosen predictor.
- Wrapper methods use machine learning algorithms to select features with the most predictive power.
- Embedded methods are used to select features during the model training phase.

For categorical features, relevance can be assessed using the chi-square score, a filter method.

Once all the pre-processing steps are completed and the data is selected, the next stage involves the modelling of the data, referred to as ***Modelling***. This phase involves selecting the algorithms that will be employed to reach the stated objective and determining how the data will be utilised within the models. These choices can be assessed through various criteria.

Following the *Modelling* phase, it is essential to evaluate the results obtained through a fifth phase known as ***Evaluation***. For this aim, a set of evaluation techniques and performance metrics were chosen.

To ensure that the results uncovered in an analysis are generalizable to an independent, unseen dataset, cross-validation is used (Larose, 2015). The most common techniques are twofold cross-validation and k-fold validation. In ***K-fold cross-validation,*** the data is split into *k* equally sized subsets, or folds. In each iteration, onefold is used as the test set, the remaining folds are used to train the model, and evaluated in the test set. This process is repeated *k* times, with each fold serving once as the test set. The model's overall performance is then estimated by averaging the evaluation results from all *k* iterations. (Kubben et al., 2019) (Figure 4)

Figure 4: K-Fold Cross Validation

Once the cross-validation procedure is defined, several error-based metrics were used to quantify the difference between predicted and actual values. In the context of this study, both *regression* and *classification models* were evaluated using appropriate metrics.

For regression tasks, which aim to predict the swimmer's velocity, the following metrics were applied:

- The ***Mean Absolute Error (MAE)*** is the mean of the absolute values of individual prediction errors across all observations (Tatachar, 2021). It can be represented by the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|,$$

where n represents the number of observations and $Y_i - \hat{Y}_i$ represents the absolute errors.

- ***Mean Squared Error (MSE)*** also measures the average magnitude of the errors, brut unlike *MAE*, it squares the individual differences, penalising larger errors more heavily (Tatachar, 2021):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

- The ***Root Mean Squared Error (RMSE)*** is the square root of *MSE*, and provides an interpretable error value in the same units as the target (Tatachar, 2021). It is given the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}.$$

- $\boldsymbol{R^2}$ or ***Coefficient of Determination*** represents the explained variance of the dependent variables by the independent variables (Tatachar, 2021). It can be represented by the following formula:

$$R^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\varepsilon^2}.$$

In contrast, for classification models, used initially in the attempted prediction of pacing strategy, the following metrics were applied to evaluate the performance:

- ***Accuracy*** measures the proportion of correct predictions made by the model out of all of the predictions (Grandini et al., 2020). It can be calculated in the following way:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)},$$

where, a *True Positive (TP)* occurs when a positive observation is correctly classified as positive by the model. A *False Positive (FP)* refers to a negative observation that is incorrectly classified as positive. A *True Negative (TN)* occurs when a negative observation is correctly classified as negative. A *False Negative (FN)* happens when a positive observation is incorrectly classified as negative by the model (Grandini et al., 2020).

- ***Precision*** measures the proportion of true positive observations among the observations that were predicted as positive by the models (Grandini et al., 2020). Can be calculated by the following formula:

$$Precision = \frac{TP}{(TP + FP)}.$$

- ***Recall*** measures the proportion of positive observations that were correctly predicted by the model out of all the actual positive observations and can be calculated by the following formula:

$$Recall = \frac{TP}{(TP + FN)}.$$

- The ***Macro F1-Score*** is a metric used to evaluate the performance of multi-call classification models and combines precision and recall (Grandini et al., 2020). It is calculated by the formula:

$$F1 - Score = 2 * \left(\frac{Macro\ Average\ Precsion * Macro\ Average\ Recall}{Macro\ Average\ Precsion^{-1} * Macro\ Average\ Recall^{-1}}\right),$$

  where Macro Average Precision and Recall are the arithmetic mean of the metric for a single class, where k is a class generic:

$$Macro\ Average\ Precision = \frac{\sum_{k=1}^{K} Precision_k}{K},$$
$$Macro\ Average\ Recall = \frac{\sum_{k=1}^{K} Recall_k}{K}.$$

Beyond standard performance evaluation, a set of statistical tests were applied to validate the assumptions of the models. These tests were used depending on the nature of the data and the modelling objective, whether classification or regression.

When building a regression model, one of the important assumptions is that the residuals (or errors) are independent from one observation to the next. If this assumption is violated, meaning the residuals are correlated, it may indicate that the model hasn't fully captured the structure of the data. This can affect the accuracy of the model's predictions and the reliability of any conclusions drawn. To assess whether this assumption holds, the ***Durbin-Watson test*** was applied, which is specifically designed to detect autocorrelation in the residuals. This test is especially useful when working with sequential or time-related data, where values close together may be more similar than expected (Hyndman & Athanasopoulos, 2018). Using the *Durbin–Watson test* helps ensure that the model is well specified and that the residuals behave as expected, random, uncorrelated, and centred around zero. If significant autocorrelation were found, it would suggest the presence of missing variables or underlying patterns not accounted for by the model.

In addition, an ***Analysis of Variance (ANOVA***) was conducted to determine whether significant differences existed between the means of various groups. This technique is a parametric statistical model used to compare the means of three or more independent groups, with the aim of identifying whether at least one of the group mean significantly differs from the others. The core principle of *ANOVA* lies in decomposing the total variance observed in the data into two components: between-group variance and within-group variance (also known as residual variance) (Afonso & Nunes, 2019). The test statistic follows the *F-distribution* and is calculated as:

$$F = \frac{MTS}{MSE},$$

Where $MTS = \frac{SSB}{K-1}$, $MSE = \frac{SSE}{n-K}$. Here, $SSB = \sum_{i=1}^{K} n_i(\bar{X}_I - \bar{\bar{X}})^2$ represents the between-groups sum of squares, and $SSE = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2$, denotes the within-group (or residual) sum of squares. K is the number of groups, and n is the total number of observations.

The null hypothesis (H_0) states that all group means are equal. Rejecting the null hypothesis in the *ANOVA F-test* only indicates that at least one group mean differs from the others, it does not specify which means are significantly different. To determine which specific pairs of group means differ, a post hoc test is required (Afonso & Nunes, 2019).

The ***Chi-Square*** test is a non-parametric statistical test used to determine if categorical variables are independent of each other. It compares the observed frequencies (O_i) of each category combination to the expected frequencies (E_i) under the assumption that the variables are independent (Afonso & Nunes, 2019). The test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i},$$

where K is the number of classes. The null hypothesis (H_0) assumes that two variables are independent. It is rejected if the test statistic exceeds the critical value for a given significance level α, normally $\alpha = 0.05$. (Afonso & Nunes, 2019)

The final phase is ***Deployment,*** in which the created model is put into action as a software or a concluding report, along with its maintenance.

4 Results

4.1 Classification of Pacing Strategies

The initial approach to this study began with a classification objective, where the goal was to predict each athlete's pacing strategy based on race characteristics. For analysing the 800m, the data was grouped by athlete, event, and race type. The variables chosen *were Entry Time, Reaction, Start Speed, Middle Speed, End Speed, CV (%), Final Sprint, Final Rank, Age, Sex, Type, and Country.*

The first step was using *FAMD,* a dimensionality reduction technique appropriate for datasets containing both numerical and categorical variables. The analysis of the screen plot pointed to the retention of two components, revealing that the first two components explained approximately 31% of the total variance, suggesting moderate structure but limited dimensionality capture (Annex 1).

Following this, *agglomerative hierarchical clustering* was applied to the two-dimensional component space using *Ward's method* and *Euclidean distance*. Although the scree plot initially suggested a solution of five clusters, this configuration did not yield clearly separable groups (Annex 2). A three-cluster solution was ultimately chosen, as it provided better separation and interpretability of pacing patterns. These clusters were manually interpreted based on speed profiles and labelled as follows:

- Cluster 1: *Slower U-Shaped*;
- Cluster 2: *Faster U-Shaped*;
- Cluster 3: *Positive Split (with slight drop at the end).*

After clustering, the clusters were treated as pseudo-labels for classification purposes, transforming the unsupervised problem into a supervised classification task. This allowed exploration of which swimmer features were most influential in predicting pacing strategy. To identify the most relevant predictors for pacing strategy classification, a combination of filter, embedded, and wrapper methods was employed. Categorical variables were evaluated using the *Chi-square test*, while numerical features were first filtered based on variance thresholds, then assessed through decision tree–based feature importance, and finally refined using *Recursive Feature Elimination (RFE).* This multi-step selection process resulted in a final feature set comprising: *Start Speed, Middle Speed, End Speed, Final Sprint, Final Rank, CV (%), Lane, Type* and *Sex.*

Four supervised learning models were trained: *DT, RF, SVM, and MLP*. Hyperparameters were tuned via *GridSearchCV*, and *DTs* were constrained to a maximum depth of 2 to avoid overfitting. The best models were compared, and the *RF* showed the best validation performance

As seen in Annex 3 and Annex 4 the *RF* model demonstrated the strongest overall performance. It achieved the highest validation accuracy and *F1-score*, indicating a strong ability to generalise and identify patterns in swimmer pacing profiles based on the selected features. The *MLP* also performed well, achieving consistent validation results. However, its perfect fit on the training data suggests a greater risk of overfitting, which is common when working with small datasets and highly flexible models. The *SVM* offered solid and stable performance,

although it lagged slightly behind *Random Forest* in terms of predictive strength, particularly on more complex or less frequent patterns. The *DT* model, while interpretable and straightforward, exhibited the weakest validation metrics, along with the largest gap between training and validation scores, highlighting limitations in generalisation when applied to new data.

4.2 Pacing Strategy Characterisation

Due to the poor quality of the *FADM* results and to better characterise pacing strategies, a second unsupervised approach was adopted using *Gower distance-based agglomerative hierarchical clustering.*

Since the sample size is small, the dataset contains mixed data, and there is no predefined number of clusters, the most appropriate method is agglomerative hierarchical clustering. Moreover, as performance times may include extreme cases or outliers (e.g., standout performances), hierarchical methods are particularly suited for detecting such atypical patterns. The *Gower distance* was used to compute dissimilarity, as it is specifically designed to handle mixed data types. *Average linkage* was selected for the clustering process, balancing the distances within and between clusters to ensure interpretability and robustness.

For the 800m events, the scree plot (Annex 5), based on the elbow method, suggested an optimal number of three clusters. On the other hand, the silhouette score peaks at two clusters (Annex 6), but the score for the three clusters is still quite high. Choosing three clusters provides a good balance between model interpretability and capturing meaningful subgroup variation.

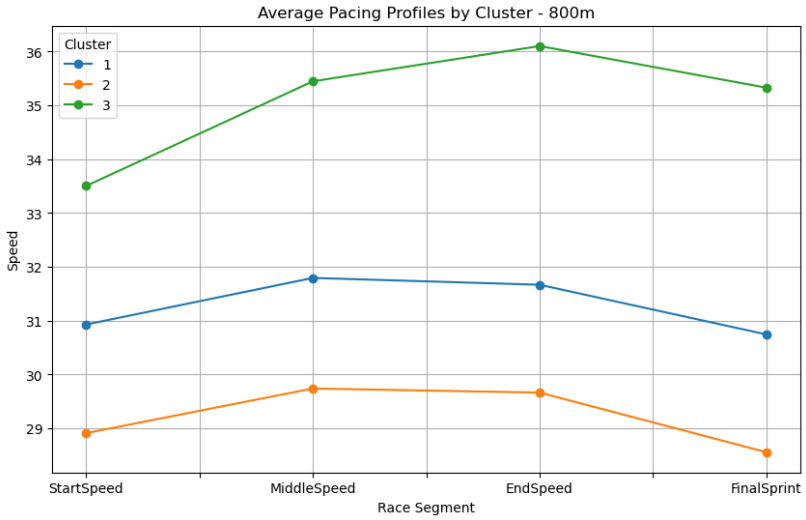


Figure 5: Average pacing profiles across race segments for each cluster in the 800m events

When analysing the pacing profiles (Figure 5), Cluster 1 and Cluster 2 followed a U-shaped strategy, characterised by faster starting and finishing segments with relatively slower middle segments. However, Cluster 2 consisted of faster athletes overall, while Cluster 1 showed similar pacing dynamics but at a slightly slower pace. Cluster 3, containing only a single athlete, followed a positive split strategy, progressively slowing down throughout the race.

To better understand the in-depth key differences, a cluster-wise comparison revealed that Cluster 1 included athletes with mid-range entry times, consistent pacing (low *CV%*), and a

higher proportion of female swimmers. While Cluster 2 was composed mostly of male athletes, with slightly higher *CV%*, suggesting more variation in pacing, but still relatively steady execution. And Cluster 3 (single athlete) had the slowest final result and erratic pacing and thus is interpreted as an atypical case (Annex 7 - Annex 12).

Now, looking into the 1500m events, two clusters emerged from the hierarchical clustering. Both demonstrated a U-shaped pacing strategy, but the main distinguishing factor was speed Cluster 1 was composed of faster swimmers compared to Cluster 2 (Figure 6).

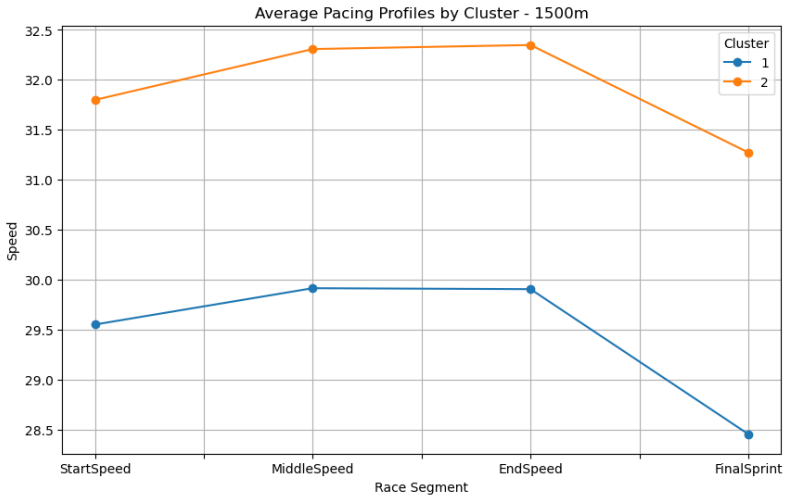


Figure 6: Average pacing profiles across race segments for each cluster in the 1500m events

Upon further analysis, Cluster 1 consisted predominantly of male athletes with faster entry times and slightly greater pacing variability and Cluster 2 included mostly female athletes, had slower entry times, but more consistent pacing, as reflected in lower *CV%*. (Annex 15, Annex 16, Annex 17, Annex 18, Annex 19 and Annex 20)

To formally assess whether the clusters differed significantly in terms of performance or demographic variables, *ANOVA* tests (for continuous variables) and chi-squared tests (for categorical variables) were conducted. The summary of results is presented in Table 3.

Table 3: ANOVA and Chi-square statistical comparison of clusters for 800m and 1500m races

Variable	Test Type	800m Results	1500m Results
Final Rank	ANOVA	$F = 1.246$ $p = 0.2950$	$F = 0.170$ $p = 0.6814$
CV (%)	ANOVA	$F = 5.070$ $p = 0.0092$	$F = 0.655$ $p = 0.4219$
Entry Time	ANOVA	$F = 171.103$ $p = 0.0000$	$F = 480.954$ $p = 0.0000$
Reaction Time	ANOVA	$F = 2.829$ $p = 0.0670$	$F = 1.246$ $p = 0.2936$
Race Type	Chi-squared	$\chi^2 = 16.107$ $p = 0.0409$	$\chi^2 = 7.472$ $p = 0.1129$
Sex	Chi-squared	$\chi^2 = 63.000$ $p = 0.0000$	$\chi^2 = 48.234$ $p = 0.0000$
Country	Chi-squared	$\chi^2 = 44.479$ $p = 0.8189$	$\chi^2 = 24.303$ $p = 0.3316$

The statistical analysis came to confirm the prior conclusions about the cluster differences it revealed that, independent of the cluster, it did not significantly explain athletes' final rankings in either the 800m or 1500m events, suggesting that while pacing strategies may influence race dynamics, they do not directly determine performance outcomes. In the 800m, there were significant differences between clusters in pacing variability (*CV%*), entry time, and race type, indicating a more tactically diverse race structure. In contrast, the 1500m showed fewer between-cluster differences, with only entry time emerging as a significant factor, suggesting a more uniform strategic approach in longer-distance events. Entry time (seed time) was the most consistent differentiator across both events, implying that pre-race expectations and physiological capacity may shape pacing behaviours. Additionally, sex was strongly associated with cluster membership in both races, reflecting gender-based differences in pacing profiles. Reaction time showed no significant differences across clusters, and nationality was not associated with the pacing cluster in either event. These results highlight that clustering primarily reflects different pacing strategies and demographic traits rather than determining competitive success.

4.3 Key Determinants of Velocity

Another major objective of this study was to identify which features most strongly influence swimming speed during the 800m and 1500m Olympic events. To explore this, a series of new performance-derived variables were computed: *Speed*, *Previous Speed*, *Previous Split*, *Acceleration*, and *Meters to Finish*. These were designed to reflect both performance status and race progression.

As a first step, *Spearman correlation* analysis (Annex 21) was applied to detect potential multicollinearity among the candidate variables. For both race distances, variables that were highly correlated were filtered to avoid redundancy and overfitting. Only one variable from

each highly correlated pair was retained to ensure the robustness of the subsequent model. The final set of predictors selected to explain velocity included: *Acceleration, Entry Time, CV (%), Distance, Age, Reaction Time, Rank, Sex, and Type*.

For this part of the study, a *Random Forest Regressor* was employed. *RF* is a robust, non-parametric ensemble learning method known for its ability to handle non-linear relationships, account for feature interactions, and tolerate noisy or unbalanced data without requiring strict distributional assumptions. It also provides feature importance rankings, making it particularly well-suited for exploratory modelling in complex datasets such as this.

The *RF* was applied to both the 800m and 1500m datasets to evaluate the relative importance of the selected predictors. In both cases, the model identified S*ex, Acceleration,* and *Distance* as the most impactful features in predicting instantaneous swimming velocity. (Annex 22 and Annex 23). This consistent result across both distance points to a shared set of determinants governing in-race speed: physiological attributes (reflected by *sex* and *acceleration*) and positional context within the race (*distance to finish*).

4.4 Predictive Modelling of Velocity

Based on the *Random Forest* feature importance results, variables with importance below 0.01 were considered negligible and thus excluded from the final model. This means the features selected were *Sex, Acceleration Distance and Entry Time.*

First it was done the estimation of the OLS models, for both events.

OLS Regression Results

Dep. Variable:	Velocity	R-squared:	0.709
Model:	OLS	Adj. R-squared:	0.708
Method:	Least Squares	F-statistic:	612.4
Date:	Sat, 27 Sep 2025	Prob (F-statistic):	2.15e-267
Time:	19:22:54	Log-Likelihood:	1685.5
No. Observations:	1008	AIC:	-3361.
Df Residuals:	1003	BIC:	-3336.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.5365	0.080	44.200	0.000	3.379	3.694
Acceleration	1.6603	0.101	16.439	0.000	1.462	1.859
Entry Time (s)	-0.0037	0.000	-23.409	0.000	-0.004	-0.003
Distance	-4.754e-05	6.53e-06	-7.275	0.000	-6.04e-05	-3.47e-05
Sex	-0.0044	0.006	-0.699	0.485	-0.017	0.008

Omnibus:	99.987	Durbin-Watson:	0.682
Prob(Omnibus):	0.000	Jarque-Bera (JB):	424.583
Skew:	0.367	Prob(JB):	6.35e-93
Kurtosis:	6.094	Cond. No.	4.77e+04

Figure 7: OLS - 800m

OLS Regression Results

Dep. Variable:	Velocity	R-squared:	0.786
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	1541.
Date:	Sat, 27 Sep 2025	Prob (F-statistic):	0.00
Time:	16:33:18	Log-Likelihood:	3242.9
No. Observations:	1679	AIC:	-6476.
Df Residuals:	1674	BIC:	-6449.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.4795	0.059	59.417	0.000	3.365	3.594
Acceleration	1.3589	0.076	17.766	0.000	1.209	1.509
Entry Time (s)	-0.0019	6.04e-05	-31.834	0.000	-0.002	-0.002
Distance	-1.809e-05	2.02e-06	-8.954	0.000	-2.21e-05	-1.41e-05
Sex	-0.0051	0.005	-1.136	0.256	-0.014	0.004

Omnibus:	534.676	Durbin-Watson:	0.739
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4188.228
Skew:	1.269	Prob(JB):	0.00
Kurtosis:	10.309	Cond. No.	1.13e+05

Figure 8: OLS - 1500m

For the estimations of the remaining algorithms, *GridsearchCV* was used, using scikit-learns module (Albon, 2018). The results from this optimal search can be observed in Annex 24 and Annex 25. To ensure the robustness and generalizability of the models, a cross-validation approach was employed. Specifically, *k-fold cross-validation* (with k = 5) was used during hyperparameter tuning to evaluate model performance across different data partitions. This helped mitigate overfitting and provided more stable estimates of model accuracy and predictive power, particularly important given the relatively small sample size of the dataset (Reyaz et al., 2022).

To identify the model with the best predictive power, the tables bellow summarize the evaluation metrics of each model for each race.

Table 4: Evaluation Metrics of the algorithms - 800m races

Algorithm	MAE	MSE	RMSE	R^2
OLS	0.032	0.002	0.046	0.709
Randon Forest	0.010	0.000	0.015	0.968
MLP	0.025	0.001	0.035	0.819
AdaBoost	0.019	0.001	0.027	0.898
Gradient Boosting	0.009	0.000	0.014	0.972

Table 5: Evaluation Metrics of the algorithms - 1500m races

Algorithm	MAE	MSE	RMSE	R^2
OLS	0.024	0.001	0.035	0.786
Randon Forest	0.008	0.000	0.012	0.976
MLP	0.015	0.000	0.021	0.925
AdaBoost	0.015	0.000	0.020	0.929
Gradient Boosting	0.009	0.000	0.012	0.974

Observing Table 4Table 5, indicate that *Gradient Boosting* consistently shows the highest predictive performance across both the 800m and 1500m races. It yields the lowest error metrics (*MAE, MSE, RMSE*) and the highest R^2 values, indicating strong model fit and minimal deviation between predicted and actual values. *Random Forest* also performs very well, closely following *Gradient Boosting* in all metrics, and can be considered a reliable alternative.

In contrast, *MLP* displays signs of overfitting and poor generalisation, with notably higher errors and a lower R^2 value, indicating that it may not be well-suited for this specific dataset, probably due to its size. *OLS*, while conceptually simple and widely used, yields the weakest performance, with the lowest R^2 v*alue* of 0.709 for the 800m and 0.786 for the 1500m events, indicating that it fails to capture much of the variance in the target variable.

These results highlight the superiority of ensemble learning approaches, particularly *Gradient Boosting*, for modelling velocity based on athlete and race characteristics in middle-distance events.

The results of the *Durbin-Watson* and the mean of the residual, to assess the quality of the models, are represented in in Table 6 andTable 7.

Table 6: Precision Metrics of the algorithms - 800m races

Algorithm	Durbin-Watson Test	Mean of Residuals
OLS	0.682	-0.000000
Randon Forest	1.969	0.00060
MLP	1.504	0.00395
AdaBoost	1.310	-0.00224
Gradient Boosting	2.000	0.00001

Table 7: Precision Metrics of the algorithms - 1500m races

Algorithm	Durbin-Watson Test	Mean of Residuals
OLS	0.739	-0.000000
Randon Forest	1.941	0.00015
MLP	1.603	0.00000
AdaBoost	1.150	-0.000178
Gradient Boosting	1.987	0.00004

Table 6 and Table 7 show that the mean residuals for all models are very close to zero, indicating that the models are generally unbiased in their predictions; no model under- or over-predicted the target values. This is a good indicator of good accuracy in terms of central tendency.

The *Durbin-Watson test* was applied to assess the presence of autocorrelation in the residuals. Values close to two indicate the absence of autocorrelation, which is the desired outcome in predictive modelling. In both races, *Durbin-Watson statistics* are similar, reflecting that the models produce highly consistent predictions across both datasets. This consistence reinforces the reliability of the results. Among the models, *Gradient Boosting* and *RF* achieved the best *Durbin-Watson values*, suggesting that their residuals behave like noise, uncorrelated and random, further confirming the model's robustness.

Finally, figures Figure 9 and Figure 10 illustrate the predicted values generated by the best-performing model, *Gradient Boosting*, alongside the actual speed data throughout the race in the test dataset.

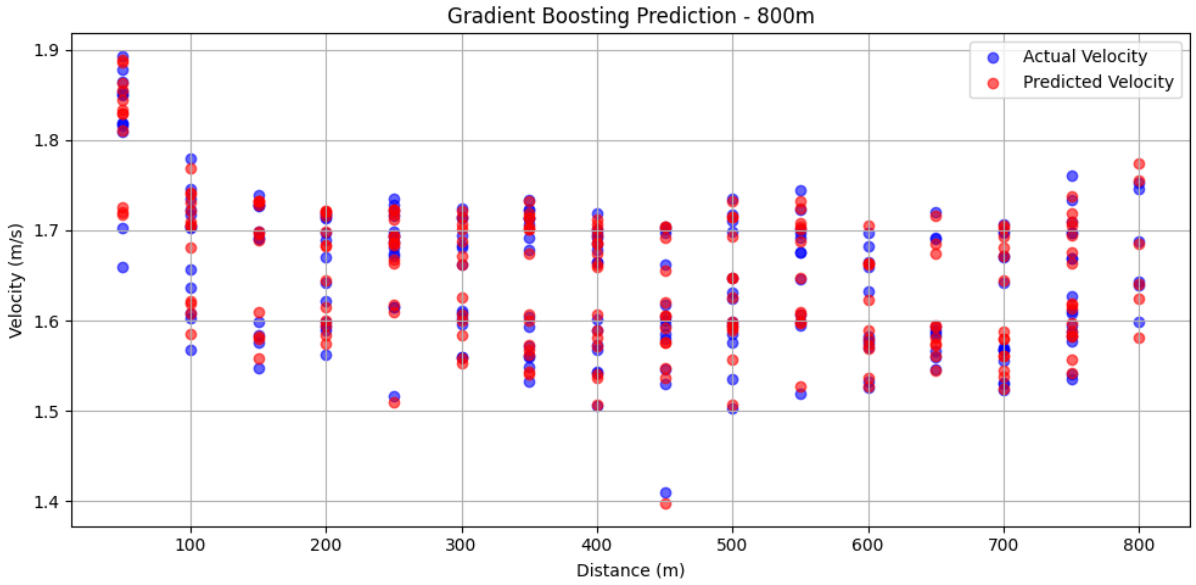


Figure 9: Predicted vs. real data of velocity across splits in the test phase for 800m events

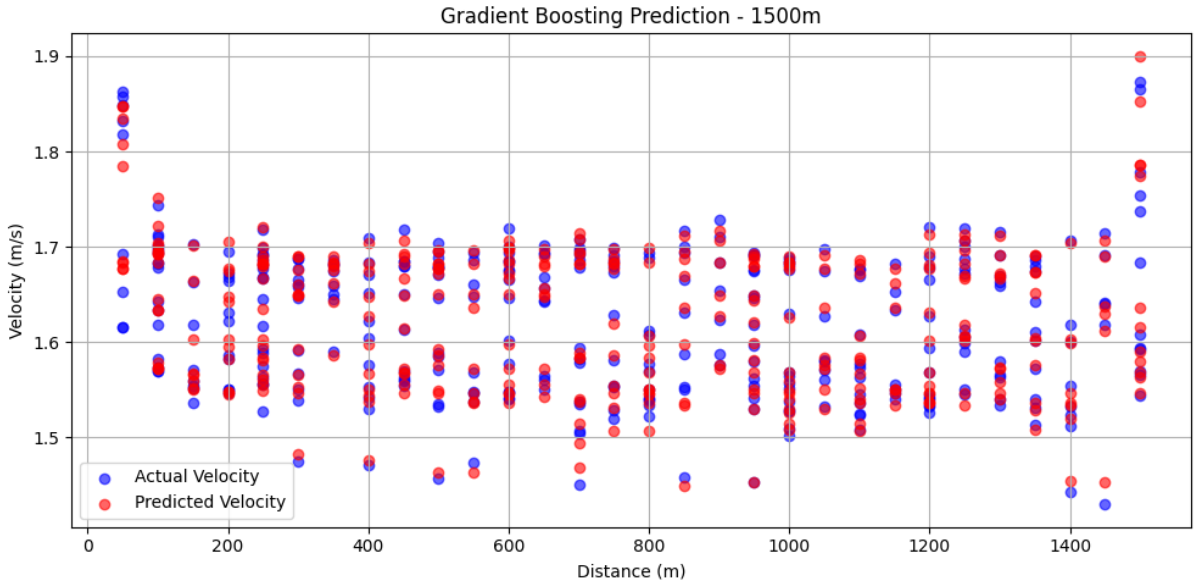


Figure 10: Predicted vs. real data of velocity across splits in the test phase for 1500m events

5 Discussion

The findings were pretty consistent with what was seen in other studies about pacing strategies in competitive swimming, especially regarding the U-shaped pacing common in long-distance events. In both the 800m and 1500m races, most swimmers followed this U-shaped profile, which aligns well with research from Morais (2023) and McGibbon (2018). These studies highlighted how elite swimmers often start and finish strong while pacing themselves more conservatively in the middle of the race to manage their energy.

In the 800m race, three different pacing profiles were spotted. Two of them followed the U-shaped path but varied in overall speed, showing a clear distinction between faster and slower swimmers. The third profile displayed a positive split, meaning there was a gradual decrease in speed throughout the race. Interestingly, demographic trends were observed: male swimmers tended to show more variability in their pacing, while female swimmers maintained greater consistency. This observation is in line with Moser (2021), who noted physiological and strategic differences between the sexes when it comes to race execution.

When looking at the 1500m, the pacing variability among athletes was less pronounced. Although U-shaped strategies remained prevalent, the primary distinction between the two groups lays in the overall speed, separation faster from slower swimmers. This relative consistency in pacing likely reflects the more significant aerobic demands of longer races, emphasising the need for careful energy management, which is in line with previous research.

Interestingly, even though there were clear pacing strategies, they didn't strongly correlate with final race rankings. This suggests that how a swimmer paces themselves isn't the sole factor in determining their success. Lara & Del Coso (2024) reflected this notion, stating that pacing is more about an athlete's physiological profile and race plan than their actual competitive placement.

In terms of what drives performance, factors like *Sex, Acceleration, and Distance* were influential across both race distances. These insights emphasise the significance of physiological traits, like power output and fatigue resistance, as well as race dynamics, in controlling speed in real-time. Entry time also played a significant role, suggesting that what swimmers achieved before the race can impact their pacing strategies during it.

On the predictive side, the *Gradient Boosting* model performed exceptionally well in terms of accuracy for swimmer velocity, with *RF* also showing strong results. Both outperformed standard methods like *OLS* and *MLP*, proving better at capturing the complex patterns and interactions in our mixed datasets. Plus, the best models showed no signs of autocorrelation in their residuals and had near-zero means, indicating their predictions are reliable. This reinforces the idea that ensemble models are excellent tools for understanding the intricacies of athletic pacing.

6 Conclusion and Future Works

This study investigated the pacing behaviour of elite swimmers in the 800m and 1500m freestyle events at the 2024 Olympic Games through a data-driven approach. By applying machine learning techniques, both unsupervised (*hierarchical clustering*) and supervised (*Gradient Boosting, Random Forests, SVM, and MLP*), it was possible to explore velocity patterns across race segments and identify the features most influential in predicting these patterns.

The results confirmed the presence of commonly observed pacing strategies, such as the U-shaped pattern, and highlighted important individual characteristics associated with in-race speed variation. While the pacing strategy itself was not explicitly predicted, the study demonstrated that certain features, such as *Sex, Event Distance, Acceleration, and CV (%)*, play a key role in shaping velocity dynamics. *Clustering* revealed consistent groupings among swimmers, particularly differentiated by sex and entry time, without a direct link to final race ranking.

Among the predictive models evaluated, *Gradient Boosting* emerged as the most accurate and robust, with low residual error and no signs of autocorrelation, as confirmed by residual analysis and the *Durbin-Watson test*. These results suggest that ensemble learning methods are highly effective in modelling complex athletic performance data and offer promising applications in sports analytics.

In practical terms, several implications emerge for athletes and coaches. The predominance of U-shaped pacing strategies suggests that training should reinforce controlled starts and finishes, alongside a steady pace through the middle race segments. The significant role of acceleration, especially in the final splits, points to the value of training that targets closing speed and fatigue resistance. Furthermore, the association between lower *CV%* and higher pacing consistency supports the inclusion of pace stability drills in elite training. The results also indicate that performance dynamics differ by sex and entry time, suggesting that training and race strategy should be tailored to individual profiles rather than following a one-size-fits-all approach.

Despite these contributions, the study has several limitations. The sample size was limited to a single competition, the 2024 Olympic Games, and certain relevant variables, such as training load, stroke efficiency, or physiological data, were not available. The pacing profiles were inferred from available velocity data, which, while useful, may not fully capture the complexity of race strategy.

Future research could build on these findings by expanding the dataset to include multiple competitions, broader demographics, or longitudinal performance data. Additionally, incorporating real-time physiological variables, such as heart rate, oxygen uptake, or lactate levels, could enhance the predictive power of the models and enable more detailed profiling of energy management and fatigue. Another direction could involve the use of wearable sensor data and real-time tracking to develop adaptive pacing tools for coaching and feedback.

References

Afonso, A., & Nunes, C. (2019). *Probabilidades e Estatística. Aplicações e Soluções em SPSS. Versão revista e aumentada*. Universidade de Évora. https://dspace.uevora.pt/rdpc/handle/10174/25959

Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions From Preprocessing to Deep Learning*.

Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, *10*(1), 5–26. https://doi.org/10.1007/s11634-014-0195-1

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bousquet, O., Luxburg, U. von, & Rätsch, G. (2004). *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*. Springer Berlin Heidelberg.

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A Methodology to Boost Data Science. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI49556.2020.9140932

Foster, C., De Koning, J. J., Hettinga, F., Lampen, J., La Clair, K. L., Dodge, C., Bobbert, M., & Porcari, J. P. (2003). Pattern of Energy Expenditure during Simulated Competition: *Medicine & Science in Sports & Exercise*, *35*(5), 826–831. https://doi.org/10.1249/01.MSS.0000065001.17658.68

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview* (No. arXiv:2008.05756). arXiv. https://doi.org/10.48550/arXiv.2008.05756

Guyon, I., & Elisseeff, A. (2003). *An Introduction to Variable and Feature Selection*.

Hołub, M., Prajzner, A., & Stanula, A. (2023). Pacing Strategy Models in 1500 m Male Freestyle Long-Course Swimming on the Basis of the All-Time Ranking. *International Journal of Environmental Research and Public Health*, *20*(6), 4809. https://doi.org/10.3390/ijerph20064809

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, *10*(5), e1380. https://doi.org/10.1002/widm.1380

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Jauhiainen, S., Äyrämö, S., Forsman, H., & Kauppi, J.-P. (2019). Talent identification in soccer using a one-class support vector machine. *International Journal of Computer Science in Sport*, *18*(3), 125–136. https://doi.org/10.2478/ijcss-2019-0021

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*(3), Article 3. https://www.informatica.si/index.php/informatica/article/view/148

Kubben, P., Dumontier, M., & Dekker, A. (Eds). (2019). *Fundamentals of Clinical Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-99713-1

Lara, B., & Del Coso, J. (2021). Pacing Strategies of 1500 m Freestyle Swimmers in the World Championships According to Their Final Position. *International Journal of*

Environmental Research and Public Health, *18*(14), 7559. https://doi.org/10.3390/ijerph18147559

Larose, D. T. (2015). *Data Mining and Predictive Analytics*.

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. B. (2019). *Supervised Machine Learning*.

Liu, P., Yuan, H., Ning, Y., Chakraborty, B., Liu, N., & Peres, M. A. (2024). A modified and weighted Gower distance-based clustering analysis for mixed type data: A simulation and empirical analyses. *BMC Medical Research Methodology*, *24*(1), 305. https://doi.org/10.1186/s12874-024-02427-8

Mahesh, B. (2020). Machine Learning Algorithms—A Review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

Markopoulou, C., Papageorgiou, G., & Tjortjis, C. (2024). Diverse Machine Learning for Forecasting Goal-Scoring Likelihood in Elite Football Leagues. *Machine Learning and Knowledge Extraction*, *6*(3), Article 3. https://doi.org/10.3390/make6030086

McGibbon, K. E., Pyne, D. B., Shephard, M. E., & Thompson, K. G. (2018). Pacing in Swimming: A Systematic Review. *Sports Medicine*, *48*(7), 1621–1633. https://doi.org/10.1007/s40279-018-0901-9

Miyamoto, S. (2022). Linkage Methods and Algorithms. In S. Miyamoto (Ed.), *Theory of Agglomerative Hierarchical Clustering* (pp. 19–42). Springer. https://doi.org/10.1007/978-981-19-0420-2_2

Morais, J. E., Barbosa, T. M., Forte, P., Bragada, J. A., Castro, F. A. D. S., & Marinho, D. A. (2023). Stability analysis and prediction of pacing in elite 1500 m freestyle male swimmers. *Sports Biomechanics*, *22*(11), 1496–1513. https://doi.org/10.1080/14763141.2020.1810749

Moser, C., Sousa, C. V., Olher, R. R., Hill, L., Nikolaidis, P. T., Rosemann, T., & Knechtle, B. (2021). Pacing in World-Class Age Group Swimmers in 200 and 400 m Individual Medley. *Frontiers in Physiology*, *11*, 629738. https://doi.org/10.3389/fphys.2020.629738

Oliveira, G. T. D., Werneck, F. Z., Coelho, E. F., Simim, M. A. D. M., Penna, E. M., & Ferreira, R. M. (2019). What pacing strategy 800m and 1500m swimmers use? *Revista Brasileira de Cineantropometria & Desempenho Humano*, *21*, e59851. https://doi.org/10.1590/1980-0037.2019v21e59851

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R* (1st edn). Chapman and Hall/CRC. https://doi.org/10.1201/b17700

Reyaz, N., Ahamad, G., Khan, N. J., & Naseem, M. (2022). Machine Learning in Sports Talent Identification: A Systematic Review. *2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, 1–6. https://doi.org/10.1109/ICEFEET51821.2022.9848247

Rokach, L., & Maimon, O. (2005). Clustering Methods. In O. Maimon & L. Rokach (Eds), *Data Mining and Knowledge Discovery Handbook* (pp. 321–352). Springer US. https://doi.org/10.1007/0-387-25465-X_15

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D'Hondt, E., & Witvrouw, E. (2020). A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Medicine & Science in Sports & Exercise*, *52*(8), 1745–1751. https://doi.org/10.1249/MSS.0000000000002305

Russell, S. J., & Norvig, P. (with Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J.). (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

Sah, S. (2020). *Machine Learning: A Review of Learning Types*. MATHEMATICS & COMPUTER SCIENCE. https://doi.org/10.20944/preprints202007.0230.v1

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315. https://ieeexplore.ieee.org/abstract/document/7724478

Tatachar, A. V. (2021). *Comparative Assessment of Regression Models Based On Model Evaluation Metrics*. *08*(09).

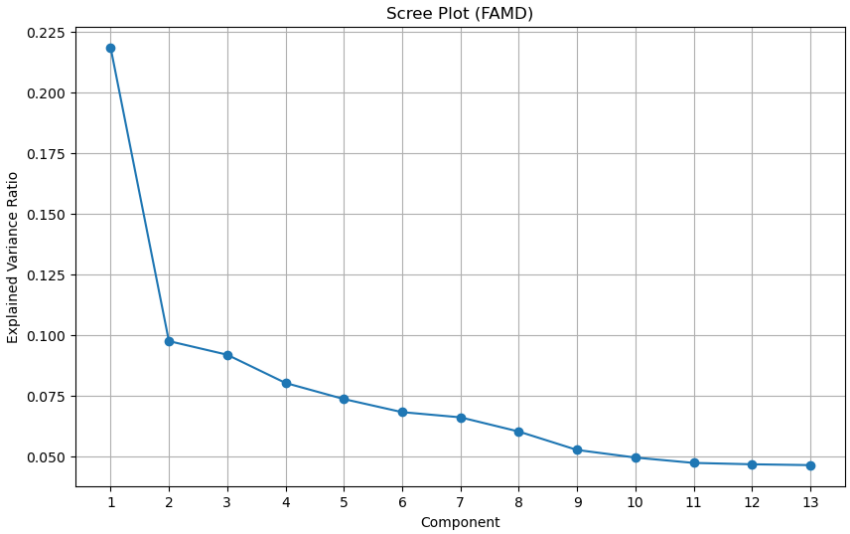
Vec, V., Tomažič, S., Kos, A., & Umek, A. (2024). Trends in real-time artificial intelligence methods in sports: A systematic review. *Journal of Big Data*, *11*(1), 148. https://doi.org/10.1186/s40537-024-01026-0

Vujovic, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/ijacsa.2021.0120670

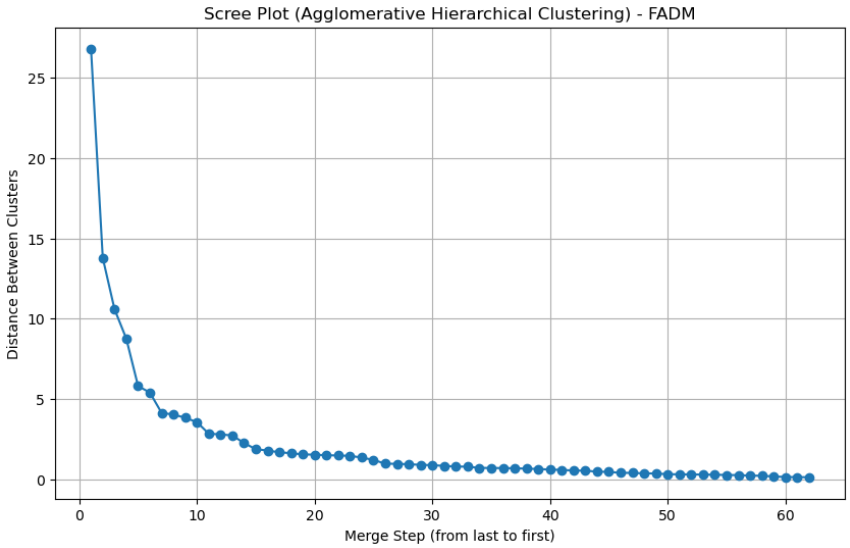
Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, *598*, 126266. https://doi.org/10.1016/j.jhydrol.2021.126266

Annexes

Annex 1: Screen plot FADM for Classification - 800m events



Annex 2: Screen plot for Agglomerative Hierarchical Clustering for Classification problem - 800m



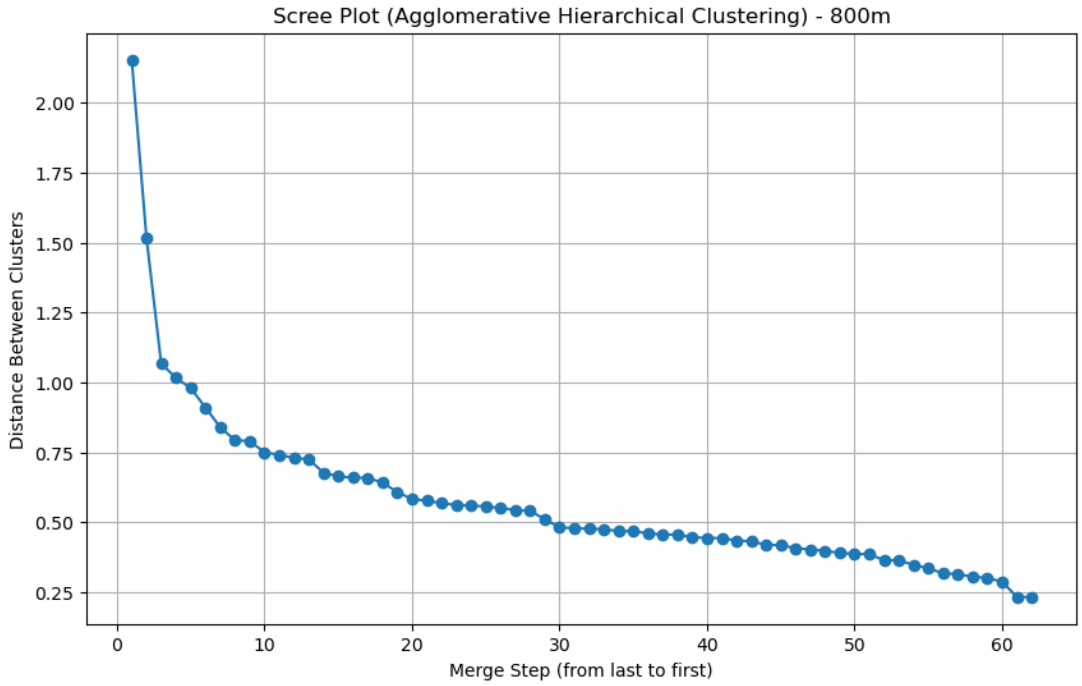
Annex 3: Model's performance for Classification problem - 800m events

	Train	Validation
Best RF	1.0+/-0.0	0.906+/-0.07
Best SVM	0.997+/-0.01	0.862+/-0.04
Best DT	0.947+/-0.02	0.802+/-0.11
Best NN	1.0+/-0.0	0.881+/-0.08

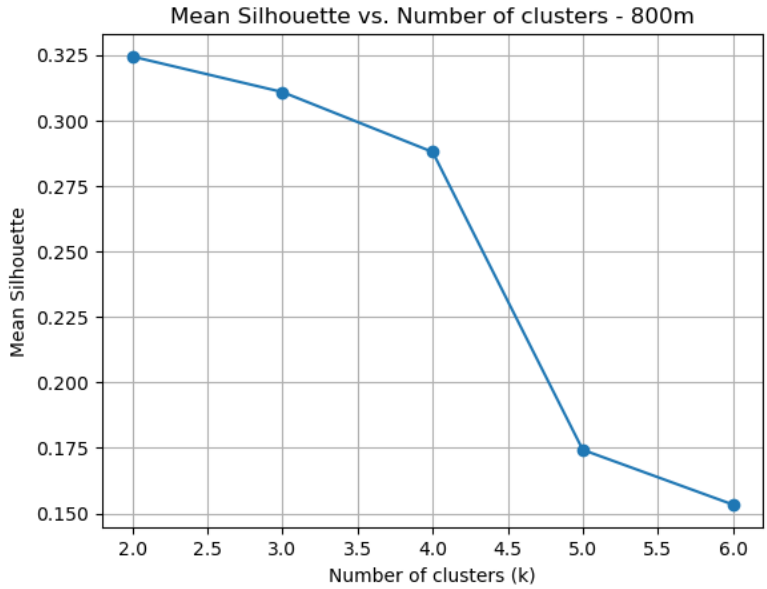
Annex 4: Evaluation metrics for the Classification problem - 800m events

```
 Random Forest
F1: 0.906 ± 0.072
Accuracy: 0.922 ± 0.049

 SVM
F1: 0.862 ± 0.042
Accuracy: 0.873 ± 0.039

 Decision Tree
F1: 0.802 ± 0.112
Accuracy: 0.827 ± 0.093

 Neural Network
F1: 0.881 ± 0.080
Accuracy: 0.891 ± 0.078
```

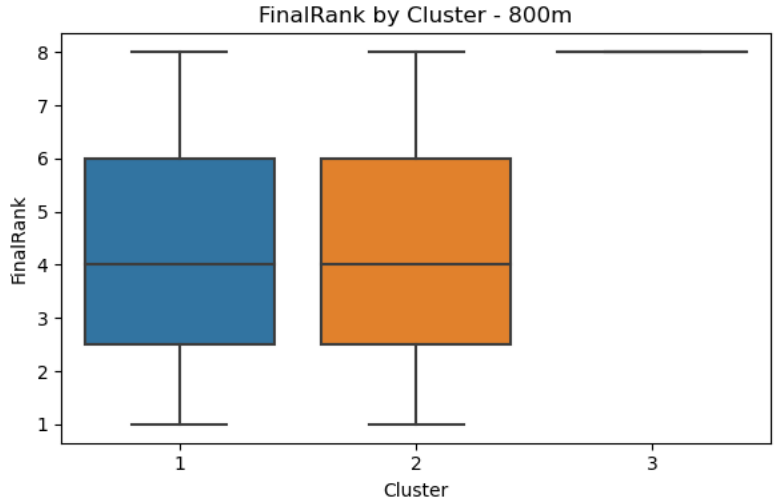
Annex 5: Screen plot for Agglomerative Hierarchical Clustering - 800m



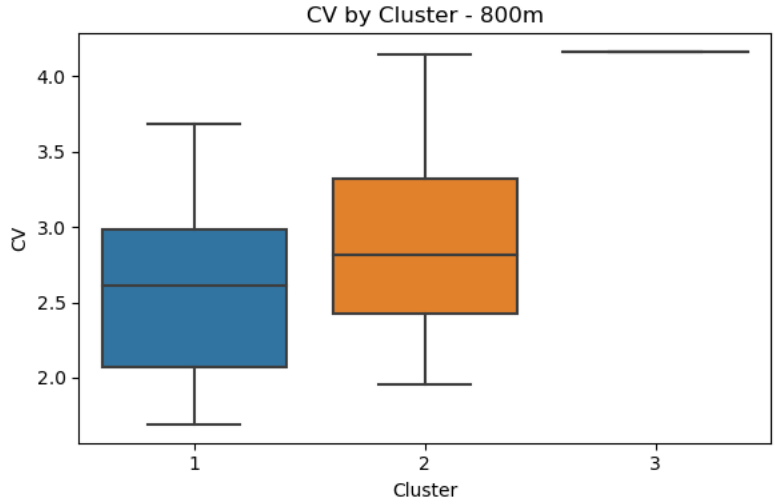
Annex 6: Silhouette scores - 800m



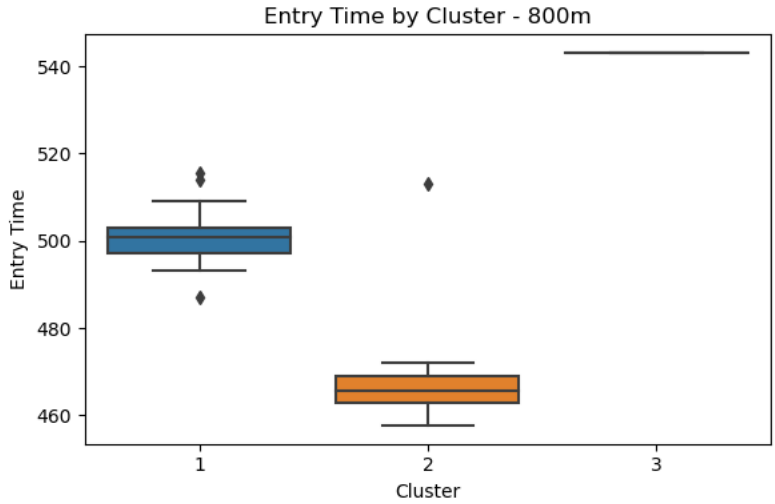
Annex 7: Boxplot of Final rank by Cluster - 800m events



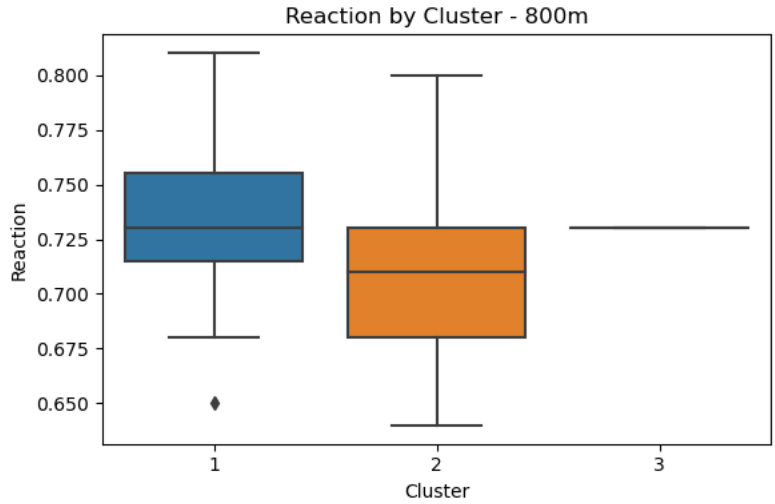
Annex 8: Boxplot of CV (%) by Cluster - 800m events



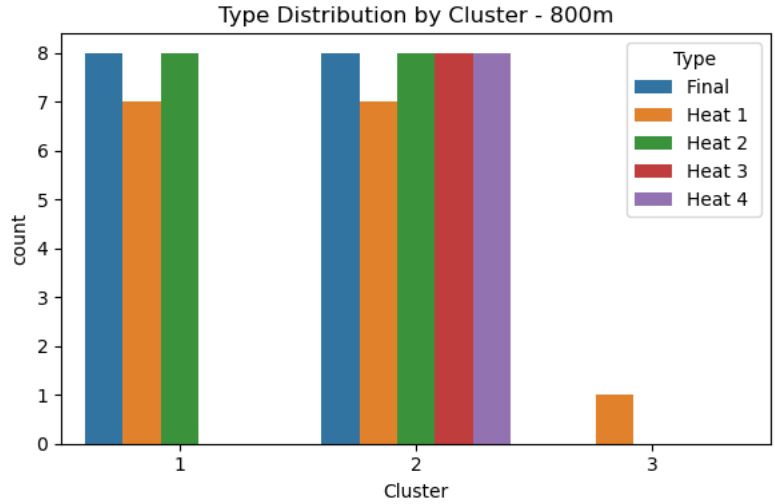
Annex 9: Boxplot of Entry Time by Cluster - 800m events



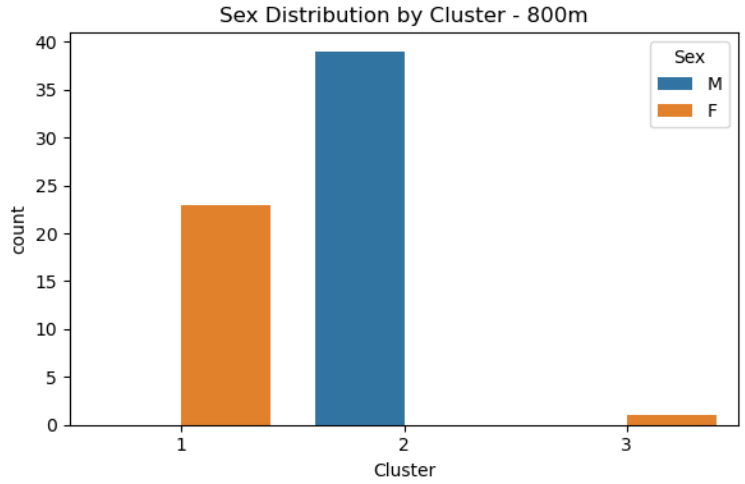
Annex 10: Boxplot of Reaction Time by Cluster - 800m events



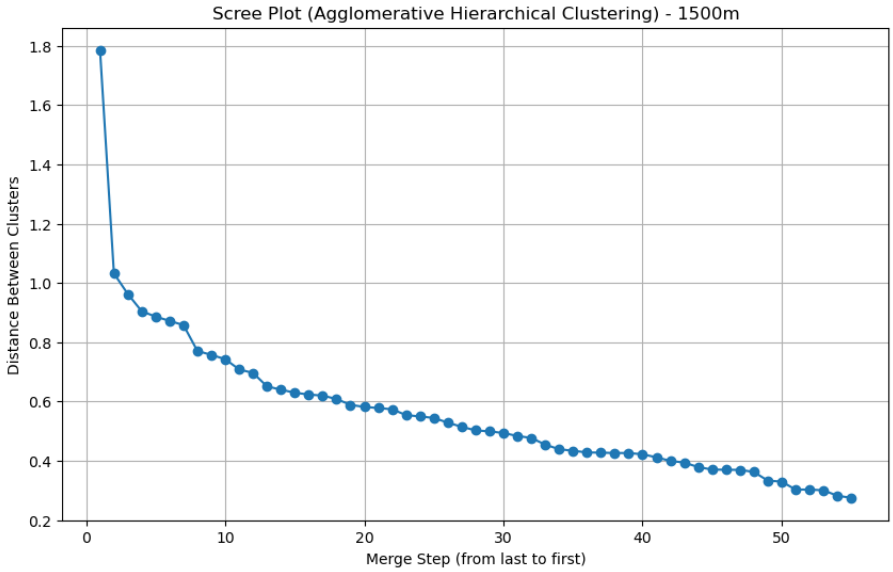
Annex 11: Bar chart of Type of race distribution by Cluster - 800m events



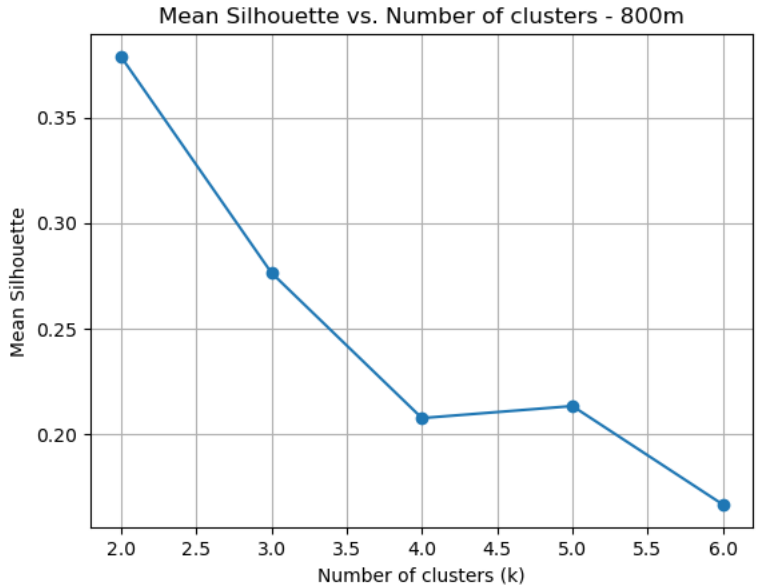
Annex 12: Bar chart of Sex distribution by Cluster - 800m events



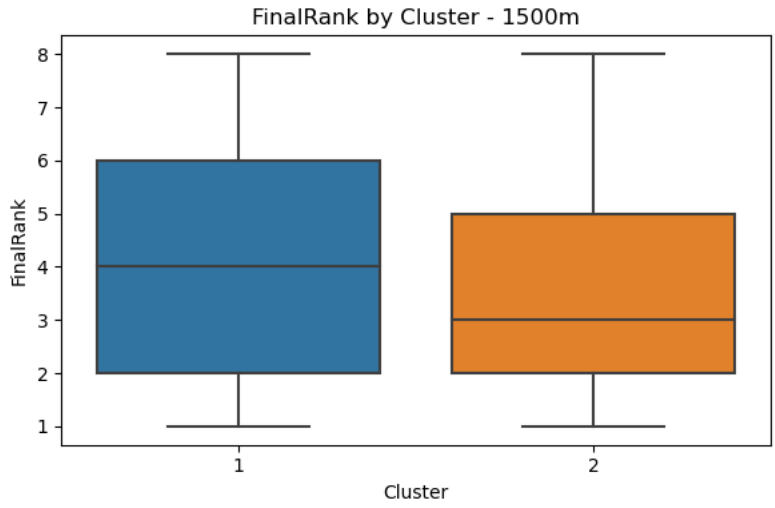
Annex 13: Screen plot 1500m



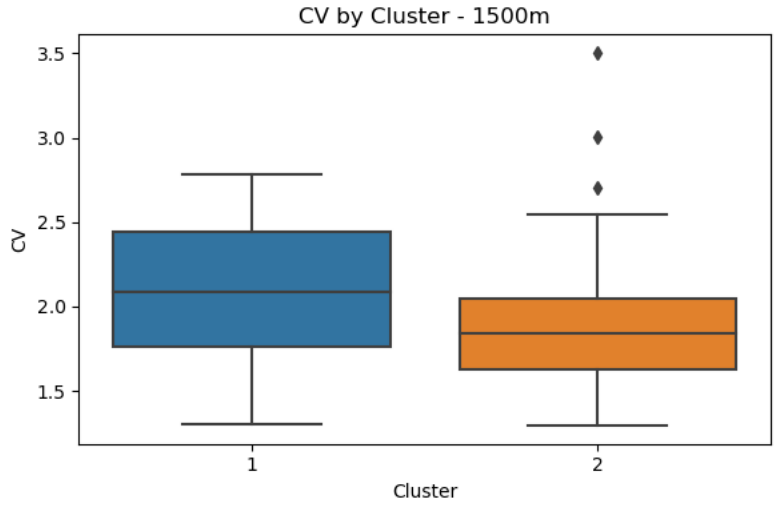
Annex 14: Silhouette scores - 1500m events



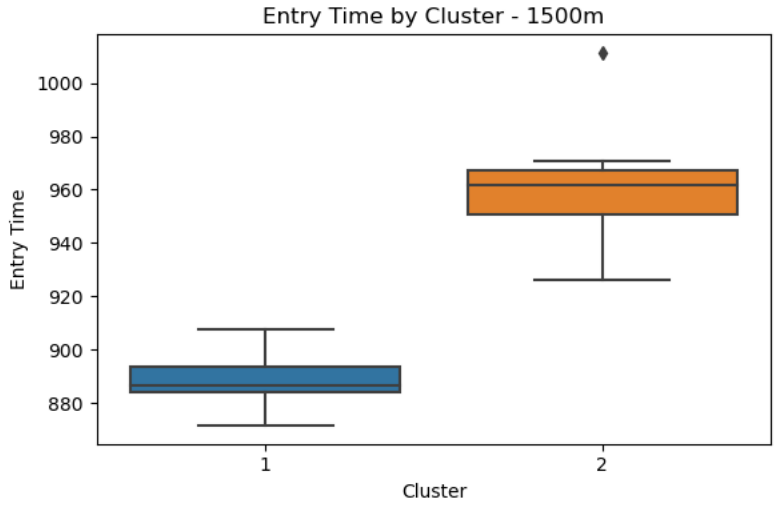
Annex 15: Boxplot of Type of race by Cluster - 1500m events



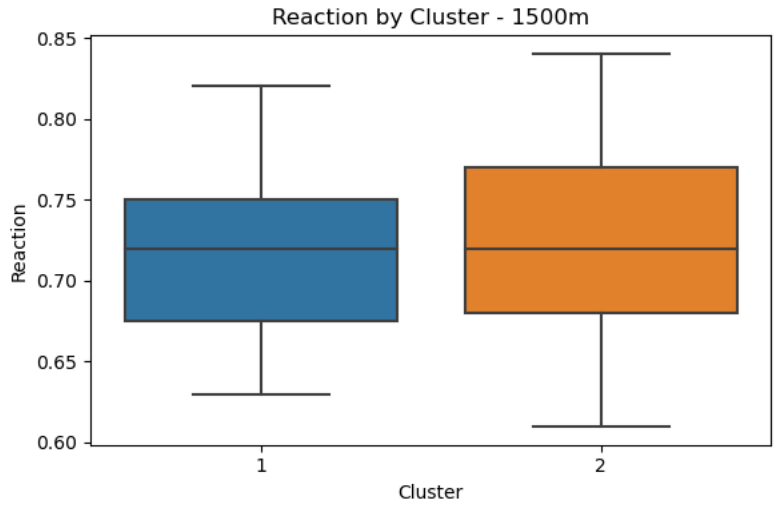
Annex 16: Boxplot of CV (%) by Cluster - 1500m events



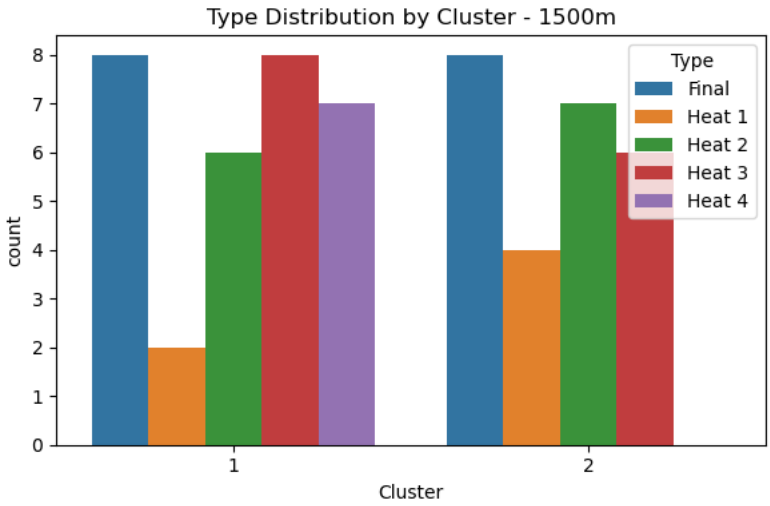
Annex 17: Boxplot of Entry time by Cluster - 1500m events



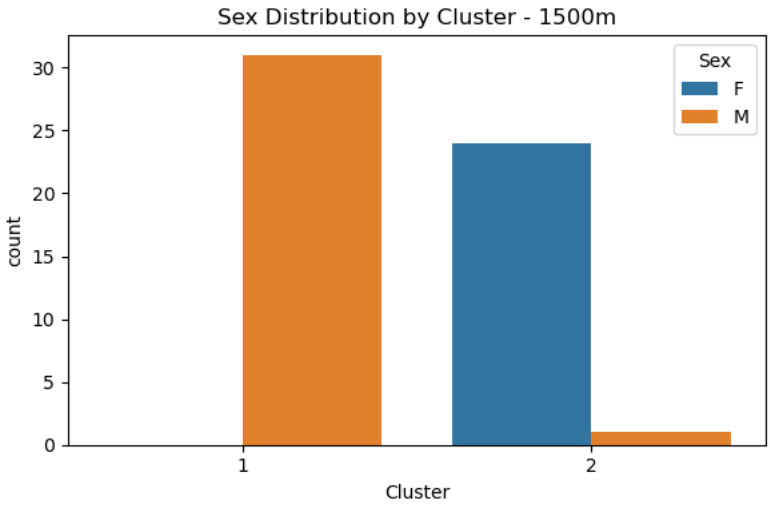
Annex 18: Boxplot of Reaction time by Cluster - 1500m events



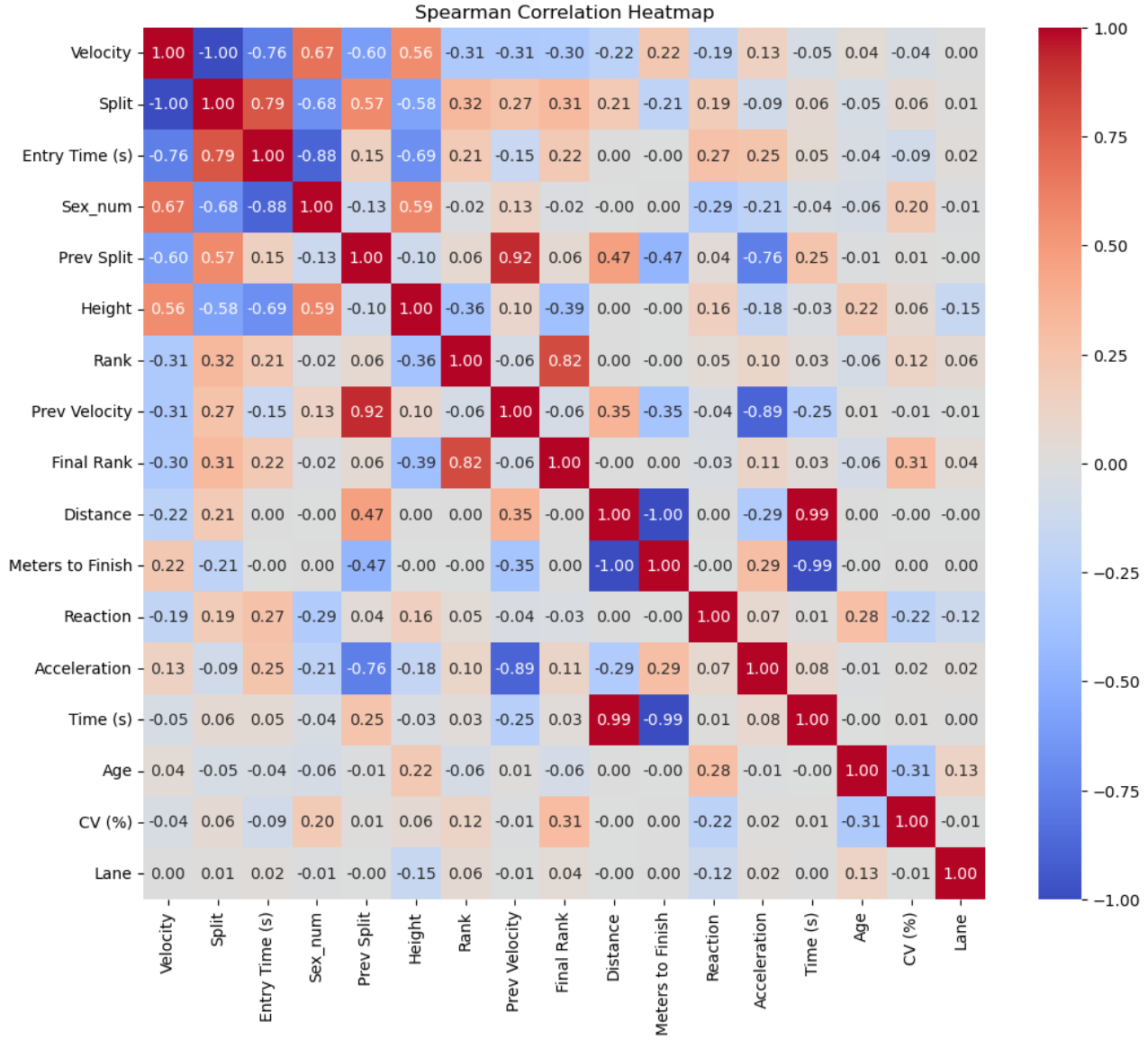
Annex 19: Bar chart of Type of race distribution by Cluster - 1500m events



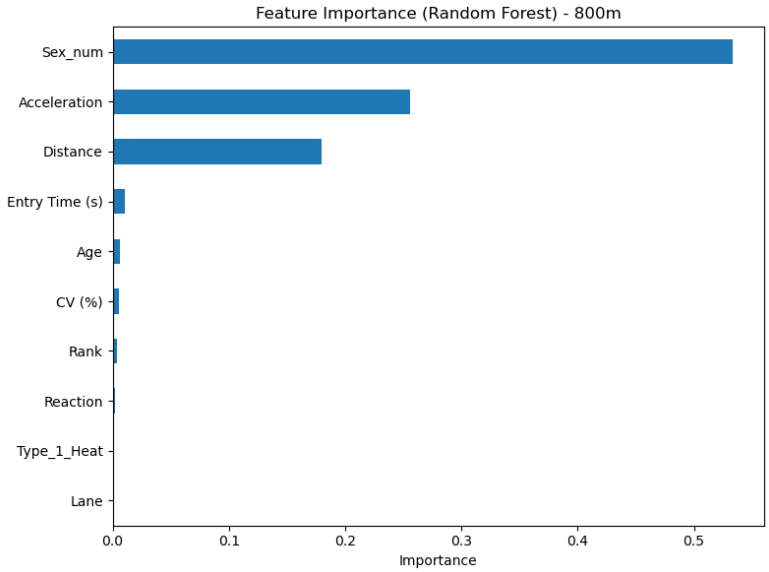
Annex 20: Bar chart of Sex distribution by Cluster - 1500m events



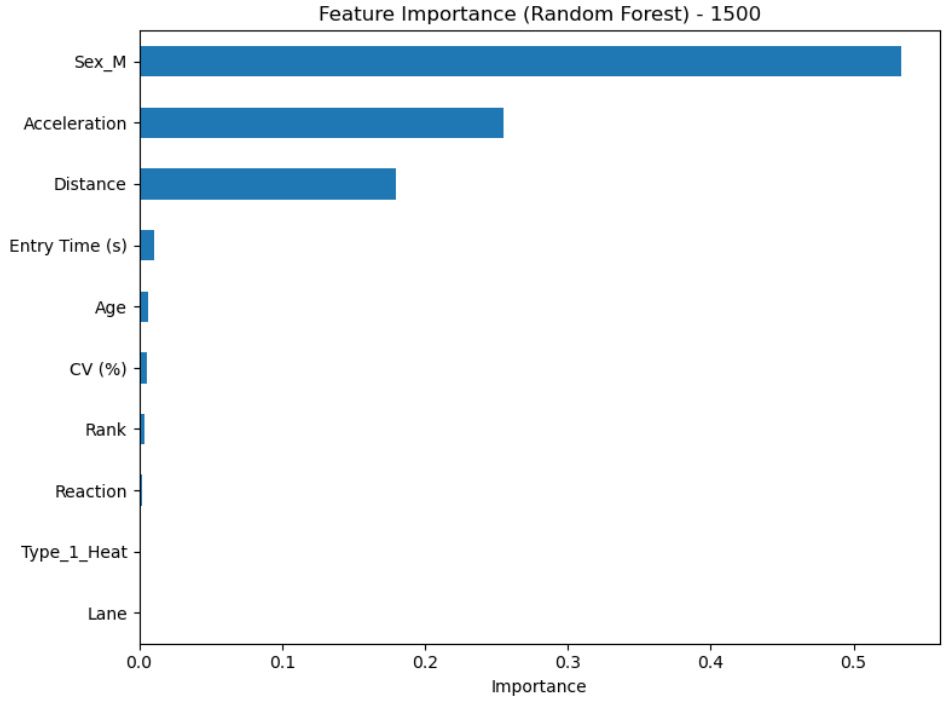
Annex 21: Spearman Correlation Heatmap for all of the races



Annex 22: Feature Importance - 800m



Annex 23: Feature Importance - 1500m



Annex 24: Best hyperparameters for ML Models using GridSearchCV - 800m events

```
 Tuning Random Forest...
 Best Params: {'regressor': RandomForestRegressor(random_state=42), 'regressor__max_depth': 10, 'regressor__n_estimators': 100}
Best MSE (CV): 0.0002

 Tuning MLP...
 Best Params: {'regressor': MLPRegressor(max_iter=1000, random_state=42), 'regressor__alpha': 0.01, 'regressor__hidden_layer_sizes': (50, 50)}
Best MSE (CV): 0.0008

 Tuning AdaBoost...
 Best Params: {'regressor': AdaBoostRegressor(random_state=42), 'regressor__learning_rate': 1.0, 'regressor__n_estimators': 50}
Best MSE (CV): 0.0006

 Tuning GradientBoosting...
 Best Params: {'regressor': GradientBoostingRegressor(random_state=42), 'regressor__learning_rate': 0.1, 'regressor__max_depth': 3, 'regressor__n_estimators': 200}
Best MSE (CV): 0.0002
```

Annex 25: Best hyperparameters for ML Models using GridSearchCV - 1500m events

```
Best Params: {'regressor': RandomForestRegressor(random_state=42), 'regressor__max_depth': 10, 'regressor__n_estimators': 100}
Best MSE (CV): 0.0002

Tuning MLP...
Best Params: {'regressor': MLPRegressor(max_iter=1000, random_state=42), 'regressor__alpha': 0.01, 'regressor__hidden_layer_sizes': (50, 50)}
Best MSE (CV): 0.0008

Tuning AdaBoost...
Best Params: {'regressor': AdaBoostRegressor(random_state=42), 'regressor__learning_rate': 1.0, 'regressor__n_estimators': 50}
Best MSE (CV): 0.0006

Tuning GradientBoosting...
Best Params: {'regressor': GradientBoostingRegressor(random_state=42), 'regressor__learning_rate': 0.1, 'regressor__max_depth': 3, 'regressor__n_estimators': 200}
Best MSE (CV): 0.0002
```