Chair of Strategy and Organization
TUM School of Management
Technical University of Munich

Predicting Performance in Competitive Swimming: An Analysis of Machine Learning Methods for Young Male Athletes

Vorhersage der Leistung im Wettkampfschwimmen: Eine Analyse von Machine-Learning-Methoden für männliche Nachwuchssportler

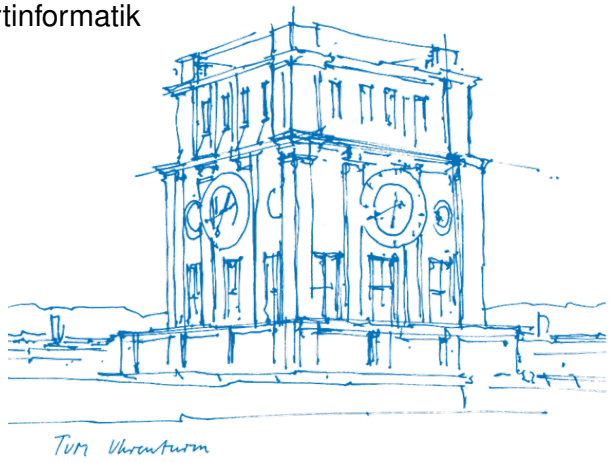
Using Random Forest, KNN, and Logistic Regression to Forecast Athletic Success

Thesis for the Attainment of the Degree
Bachelor of Science
at TUM School of Medicine and Health

Examiner
Dr. Tiago Russomanno
Lehrstuhl für Trainingswissenschaft und Sportinformatik

Submitted by
Felix-Daniel Bongarz
Guido-Schneble Str. 40, 80689 München
Matriculation Number: 02885545

Submitted on
30.09.2024



Abstract

This bachelor thesis investigates the utilization of machine learning models to forecast the performance of young swimmers by analyzing historical competition data. The main objective is to determine if a swimmer can achieve 750 FINA points by the time they reach 18 years old. In this analysis, three different models were utilized: Random Forest, K-Nearest Neighbor, and Logistic Regression. The dataset comprises competition results for male swimmers aged 11 to 18 years, who ranked among the top 100 from 2001 to 2023. To validate the models' reliability, a five-fold cross-validation method was conducted.

The findings indicate that the random forest model yielded the most precise predictions, achieving an AUC of 0.97 and an accuracy rate of 92.97%. Key predictors included FINA scores at age 16, yearly performance enhancements, and the youngest age at which an athlete broke into the top 100. Conversely, while the KNN model demonstrated high accuracy, it exhibited a reduced detection rate of pertinent cases, and logistic regression performed below average due to the model's linear characteristics.

This research offers significant insights into the factors that lead to long-term success in swimming and underscores the potential of machine learning methods in forecasting athletic performance. The outcomes provide practical recommendations for developing personalized training regimens and recognizing talent at an early stage.
Keywords: *Machine Learning, Swimming,Age Group, Prediction.*

Abstract (german)

Diese Bachelorarbeit untersucht die Anwendung von Modellen des maschinellen Lernens zur Prognose der Leistung junger Schwimmer anhand historischer Wettkampfdaten. Das primäre Ziel besteht darin, vorherzusagen, ob ein Schwimmer bis zum Alter von 18 Jahren 750 FINA-Punkte sammeln kann. Für diese Analyse wurden drei verschiedene Modelle verwendet: Random Forest (RF), K-Nearest Neighbor (KNN) und Logistic Regression. Der Datensatz umfasst Wettkampfergebnisse für männliche Schwimmer im Alter von 11 bis 18 Jahren, die von 2001 bis 2023 zu den Top 100 gehörten. Um die Zuverlässigkeit der Modelle sicherzustellen, wurde eine fünffache Kreuzvalidierung durchgeführt.

Die Ergebnisse deuten darauf hin, dass das Random-Forest-Modell mit einer AUC von 0,97 und einer Genauigkeitsrate von 92,97% die genauesten Vorhersagen lieferte. Zu den signifikanten Prädiktoren gehörten FINA-Werte im Alter von 16 Jahren, jährliche Leistungsverbesserungen und das früheste Alter, in dem ein Athlet in die Top 100 kam. Im Gegensatz dazu wies das KNN-Modell eine hohe Genauigkeit auf, wies jedoch eine geringere Erkennungsrate relevanter Fälle auf, während die logistische Regression aufgrund der linearen Eigenschaften des Modells unterdurchschnittlich abschnitt.

Diese Studie bietet wertvolle Einblicke in die Elemente, die zum langfristigen Erfolg beim Schwimmen beitragen, und unterstreicht das Potenzial von Techniken des maschinellen Lernens bei der Vorhersage sportlicher Leistungen. Die Ergebnisse bieten praktische Anleitungen für die Erstellung individueller Trainingsprogramme und die frühzeitige Erkennung von Talenten.

Keywords: *Machine Learning, Leistungsschwimmen, Nachwuchssportler, Prediction.*

Contents

List of Tables

List of Figures

1 Introduction

With the rise of digital technology, the significance of Artificial Intelligence (AI), Machine Learning (ML), and Big Data has grown across many sectors, including sports. In swimming, where fractions of a second can determine victory or defeat, Machine Learning is becoming increasingly important. This thesis explores the use of Machine Learning to predict competition results in swimming and how these technologies can be leveraged to optimize training methods and enhance competition strategies.

Success in swimming is influenced by numerous factors, including physical attributes, the social environment (family, coaches, teammates), the quality of training facilities, and the optimization of training programs. Analyzing long-term performance trends of swimmers provides valuable insights into these factors and allows for predictions of future competition outcomes. The analysis of long-term performance changes in a career based on past competition results is a useful approach for understanding and predicting athletic success at the world level (Allen & Hopkins, 2015; Allen et al., 2014; Allen et al., 2015). This type of analysis has been applied in track-and-field athletics (Haugen et al., 2018), swimming, triathlon (Malcata et al., 2014), and cross-country skiing (Walther et al., 2022). Several research projects on the subject (Berthelot et al., 2019; Walther et al., 2022) modeled performance changes and estimated the age and level of performance peaks over time. The typical empirical pattern of performance development consists of an exponential acceleration during the pubertal developmental phase, leading to a performance peak that is sustained in a plateau over 1 to 4 years. However, population-wide patterns do not take into account heterogeneity and intra-individual fluctuations (Foss et al., 2019; Haugen et al., 2018). Indeed, the analysis of individual trajectories reveals a large heterogeneity with fluctuations representing multiple periods of micro-progressions and regressions of annual individual best performances or individual positions in the world rankings throughout the entire careers of athletes (Berthelot

et al., 2011; Boccia et al., 2017; Foss et al., 2019), triathletes (Malcata et al., 2014), and swimmers (Yustres et al., 2020).

Performance improvements can differentiate the level of an athlete in the post-pubertal phase, which extends from approximately 14 years of age for girls and 16 years of age for boys. Studies in swimming (Pyne et al., 2004), athletics (Foss et al., 2019), and cross-country skiing (Walther et al., 2022) have shown that performance improvement is higher in world-class athletes than in non-world-class athletes, and in athletes ranked in the top 10 in the world compared to those ranked 11 to 100. Within-sport diversification in swimming at a younger age correlates positively with success at age 18, indicating that a higher degree of diversification is associated with higher performance level at 18. Furthermore, entering the German age group top 100 rankings at a younger age was linked to achieving higher FINA point scores at 18. This suggests that early engagement in competitive swimming may lead to better long-term performance outcomes. Conversely, early specialization in a single event, stroke, or distance category may potentially hinder long-term elite success (Staub et al., 2020). Only 33 of swimmers ranked in the German age group top 100 at 11 years were also ranked in the top 100 at 18, with 23 consistently ranked over 8 years (Staub et al., 2019). This phenomenon can be attributed to the variations in biological maturity among individuals (Abbott et al., 2021) and the influence of the relative age effect (Cobley et al., 2018). Consequently, contemporary talent development frameworks prioritize late selection to foster elite success. Similar to how the tip of a pyramid expands by widening its base, subsequent selection enhances the talent pool.

Certainly, Svendsen et al. (2018) recorded a growing number of national governing bodies that have embraced long-term development frameworks in order to implement a systematic method for training their young athletes. Nonetheless, late talent selection implies that the available resources, such as infrastructure and individual support, are

spread across a larger number of athletes (Gulbin et al., 2013), while opting for early selection would enable the concentration of resources on a smaller cohort of swimmers. Making accurate success predictions is vital to ensure that resources are allocated to the most promising swimmers rather than to transient talents who might not thrive in their final years. Given that the potential for annual performance enhancements is limited, it is improbable that swimmers whose performance significantly lags behind their age group's average will attain peak performance as they mature (Alshdokhi et al., 2020). Therefore, it is essential to establish precise performance benchmarks that can be utilized to evaluate success probabilities at their maximum age, while also ensuring that each age group acquires the foundational skills necessary to evolve into world-class adult swimmers. These performance criteria may differ based on gender (Kozieł & Malina, 2017) and specific race distances (Born et al., 2022).

Anticipating performance in swimming is crucial for enhancing training methods and devising strategies for competitions. It allows for an impartial evaluation of athlete progress and facilitates comparisons among them, thereby aiding in the focused development of talent and the identification of promising young swimmers (Allen et al., 2014; Staub et al., 2020; Staub et al., 2019). Swimming, which demands technical accuracy, stamina, and speed, necessitates that athletes and coaches thoroughly assess strengths and weaknesses, allowing for the adjustment of training programs as needed.

This bachelor thesis aims to investigate and compare the effectiveness of various machine learning models in forecasting future swimming performance. The emphasis is on developing a predictive model that identifies the key factors influencing the performance enhancement of young swimmers, along with an evaluation of different machine learning algorithms in terms of their precision and reliability. The model is anticipated to identify the key indicators of athletic achievement, with the objective variable being a performance category, like a 750-point score. The research seeks to

address the following questions: To what extent can machine learning models predict the future performance of swimmers based on historical competition data? Which machine learning models are most effective in accurately forecasting swimming performance?

2 Analysis and prediction of swimming performance: Present status of research and techniques

In this chapter, I aim to illustrate to the reader the evolution of methods used for predicting swimming performance over the years. Specifically, it emphasizes how the shift from conventional statistical techniques to contemporary machine learning approaches has impacted performance prediction strategies. The examination centers around the progression of swimming performance, the recognition of key influencing factors, and the implementation of diverse predictive models.

The advancement of sports science has highlighted the significance of modeling and predicting athlete performance, while considering elements such as training, physical condition, and gear. Unlike conventional statistical techniques, the intricate nature of performance prediction has prompted the emergence of sophisticated automated forecasting models that leverage advancements in information technology and artificial intelligence.

In the realm of swimming, analyzing and predicting competition performance is crucial, even though conventional statistical techniques have long served as the foundation for a thorough understanding of athlete development. Before the advent of machine learning and sophisticated data analysis methods, both research and training predominantly depended on traditional statistical techniques to discern performance patterns, monitor swimmers' progress, and anticipate future achievements (Allen et al., 2015; Alshdokhi et al., 2020; Born et al., 2023; Mujika et al., 2023). The academic literature presents a range of analytical and predictive strategies in swimming, encompassing both conventional statistical methods and contemporary machine learning approaches. This summary emphasizes the variety of these techniques, their applicable domains, and

the insights derived from them, thereby demonstrating the significance and potential of different analytical methods in this discipline.

2.1 General performance development and peak performance in swimming

Swimming features a wide variety of events and styles, which complicates the assessment and forecasting of swimming performance. However, there are overarching trends and patterns that can be expressed concerning performance improvement and optimal outcomes in swimming. Based on my research, numerous methodologies and approaches exist for the analysis and forecasting of swimming performance. Some of these methods utilize linear regression models to explore the connections between different variables such as training intensity, age, body composition, and experience level.

According to Staub et al. (2019), the researchers examined the performance dynamics of adolescent swimmers aged 13-18 years and found that the annual improvement in performance depends on various factors. Some of these factors include training volume, specific swimming technique, age, and competitive level. Specialization during youth also influences performance development and can have a positive impact on peak performance.

Various studies have examined the improvement in swimming performance over the years. Costa et al. (2013) observed young male swimmers over a period of two years and found that training volume had a significant influence on performance improvement. Important performance predictors for an increase were the swimming speed at which the blood lactate concentration reaches 4 mmol/L, the stroke rate, and the stroke length. Furthermore, the authors suggest that due to the unique way each swimmer adapts to training, coaches should focus on creating personalized training sessions to effectively improve an individual swimmer's performance.

It is also of great interest to examine the extent to which elite athletes actually reach the elite level. There are various approaches to this topic: Studies on performance trajectories have found that the temporal development of elite swimmers' annual personal best times is often unstable, suggesting low stability in performance improvements (Allen & Hopkins, 2015; Costa et al., 2014; Costa et al., 2011). This insight leads to the suggestion that early successes in swimming are not necessarily indicative of long-term elite performance and may potentially be avoided.

In their systematic review, Allen and Hopkins (2015) examined the age at which elite athletes achieved their peak performance, which can help determine the ideal age for talent identification. Sokolovas (2006) analyzed youth athlete rankings to study performance development and long-term athlete trajectories. They examined the presence of 17-18 year old girls and boys in the US swimming all-time top 100 rankings in relation to their earlier career successes. Many participants who were in the top 100 in younger age groups were no longer in the top 100 at ages 17-18.

Staub et al. (2019) investigated how early success and specialization in swimming impact career consistency over an 8-year period. They found that athletes who were successful early on and specialized early were less likely to maintain consistent performance over 8 years and tended to stagnate or withdraw in later career stages.

Dormehl et al. (2016) analyzed the performance of male adolescents on both regional and national scales, revealing varying rates of improvement and peak performance between the ages of 18.5 and 19.6 years. The models developed in this study can aid in talent identification by forecasting the peak age and anticipated improvement rates for swimmers. These models employed quadratic functions for both fixed and random effects to accommodate individual variations from the sample mean.

It has been suggested that a high level of junior race performance is a pre-requisite in transitioning into a successful senior athlete (Yustres et al., 2019) (Svendsen et al., 2018).

It would appear logical to track the progression of swimmers who have been successful at senior WC level in attempt to predict the pathway and requirements of swimmers for the future. Minimizing drop-out from junior to senior levels and accurate performance guidance models would be valuable in improving long term success (Costa et al., 2011) (Allen et al., 2014). It has been suggested that the most significant reason for swimming drop-out in both genders and all age groups was "dissatisfaction/other priorities (e.g., education)" Monteiro et al. (2017). If a clearer predictive model was available, swimmers, coaches, and parents would have a better guide to develop their talent and improve satisfaction within the sport based on setting realistic expectations. It is likely that this planning would improve long-term performance and improve the retention of swimmers within the competitive system.

2.2 Application of Machine Learning for Performance Prediction in Swimming

After outlining the general trends in performance development and the key factors influencing swimming, attention now shifts to contemporary techniques for predicting performance. Specifically, the examination below delves into the application of machine learning methods, which can more effectively encapsulate the intricacies and dynamics of sports performance compared to conventional statistical models.

The prediction of an athlete's performance is seen as a regression challenge, where future outcomes are estimated using past data. This field is categorized into linear and nonlinear regression approaches. Linear regressions, especially multiple linear regressions, are favored due to their straightforwardness when the factors influencing performance are few. Nevertheless, the intricate nature of sports performance, shaped

by numerous sport-specific and external elements, results in unpredictability that linear models frequently fail to address. This is where nonlinear models, specifically neural networks derived from machine learning, become relevant. These models are exceptionally adept at capturing the dynamic and unpredictable features of athletic performance.

The research conducted by the team under the leadership of Edelmann-Nusser (Hohmann et al., 2000) (Edelmann-Nusser et al., 2001) (Bügner, 2005) (Edelmann-Nusser et al., 2006) demonstrated two decades ago how performance in competitions can be anticipated from training data gathered during the preparation for immediate competitions. Their methodology relied on non-parametric modeling to address the intricate and nonlinear dynamics between training load and performance outcomes. For instance, three distinct models have been established in this regard that focus on various elements of training load: the taper phase (model T), the high-stress phase (model H), and a comprehensive model that integrates data from both phases ((Edelmann-Nusser et al., 2001), S. 20 - 21). The execution was performed using a multilayer perceptron, which analyzes the training data categorized into five different segments for their correlation with performance. During the Olympic Games in 2000 and 2004, the effectiveness of this method for forecasting swimming performance was validated. The discrepancy between the predicted and actual performance of a female swimmer in the 200-meter backstroke was merely 0.05 seconds in 2000 and 0.41 seconds in 2004. Nonetheless, the researchers noted that the validation suggested the possibility of greater margins of error (Edelmann-Nusser et al., 2001). A simulation analysis identified particular difficulties faced during the preparatory phase of German swimming teams for the 2000 Olympic Games. Logistical issues encountered while traveling to Australia resulted in unexpectedly diminished performance. The simulations indicated that improved outcomes could have been achieved under standard conditions (Edelmann-Nusser et al., 2001).

In Bügner, 2005's research, the performance of a junior swimmer over a span of 250 weeks was examined. The levels of prediction accuracy attained in earlier studies were not reached, likely due to the athlete's physical maturation. Building upon this research, Edelmann-Nusser et al., 2001 expanded their investigations to encompass a broader data set and a comparison of methodologies. The objective was to assess the effectiveness of regression and neural models in forecasting swimming performance. Data from 249 young swimmers was evaluated, incorporating fitness, strength, technique parameters, and anthropometric information. Nonlinear regression models were juxtaposed with neural networks, with the latter enabling a notably more precise prediction of swimming performance. This corroborated the superiority of neural networks in forecasting performance and identifying talent. The study also indicated that factors such as dynamic strength, swimming technique, and body mass are critical determinants of race times (Edelmann-Nusser et al., 2001).

In their dissertation, de Jesus et al., 2018 investigated backstroke start performance in swimming. They integrated both linear and nonlinear machine learning models to meticulously document the start phase in backstroke. The analysis relied on extensive data collection of swimmers competing in real conditions, incorporating kinematic and kinetic details along with biomechanical information. Neural networks (ANNs) demonstrated superior accuracy in forecasting the 5-meter start time relative to linear models, particularly by considering bounce force and immersion depth, which had notable correlations with competitive performance (Vantorre et al., 2014).

Haar, 2011 analyzed the impact of training on athletic performance by means of artificial neural networks (ANN). Over a span of three months, daily training data, encompassing heart rate readings and psychological assessments, was gathered from triathletes. This information was utilized as input variables to forecast the alterations in maximum oxygen uptake (VO2max) as an output variable. The findings indicated

that ANN demonstrated high model quality and predictive precision, emphasizing the significance of a multidimensional perspective on the training process for accurate modeling of training impacts.

Mujika et al., 2023 conducted an analysis on the performance of elite swimmers over an eight-year period to forecast swim times for the 2024 Olympic Games in Paris. They employed linear regression models alongside a Bayesian framework. The predictions, which were validated through comparisons with the outcomes of the 2022 World Championships, demonstrated a significant degree of accuracy. Notably, it was observed that American athletes frequently excelled in relay events compared to individual races. Additionally, random forest models were found to be particularly effective in forecasting medal achievements (Wu et al., 2021).

Xie et al., 2016 made a significant contribution to swimming competition prediction by systematically applying machine learning to a large dataset of over 4 million swimming records. They used various machine learning methods such as Support Vector Regression (SVR) and Random Forest (RF) to predict swimming times. The newly developed 'Wisdom of Crowd Classifier' (WoCC), which aggregates the predictions of the individual methods, showed consistently better prediction accuracies than the individual methods. The analysis revealed several age-dependent trends in swimming that provide valuable information for coaches and athletes.

The research paper authored by Tchamkerten et al., 2024 pinpointed crucial elements linked to securing Olympic swimming medals, such as nationality, level of competition, frequency of competitions, advancements, setbacks, and performance metrics prior to significant events. A logistic regression model was constructed, which attained an area under the ROC curve of approximately 90 on the test dataset. The likelihood of winning a medal rose alongside performance enhancements prior to competitions and

diminished as setbacks increased. The study highlighted the significance of advancements, minimizing setbacks, and elevating performance prior to major competitions for achieving success in international swimming.

To summarise, it can be said that machine learning methods have the potential to significantly improve the predictive accuracy of swimming performance, particularly by taking complex, non-linear relationships into account. The models introduced in this chapter provide useful methods for practical applications, particularly in talent identification and training management.

Developing a more comprehensive and accurate ML model would deepen our insight into performance variations at different stages of development and pathways to success. However, a retrospective analysis of young elite swimmers who have achieved high performance at the beginning of their adulthood and an investigation of the predictive factors of success are still pending.

This research seeks to address the following questions: To what extent can machine learning models predict the future performance of swimmers based on historical competition data? Which machine learning models are most effective in accurately forecasting swimming performance?

3 Methods

The data set was sourced from the German Swimming Association e.V. (DSV), which serves as the national governing body for swimming in Germany, through their website at www.dsv.de. We utilized the top 100 rankings of male swimmers aged 11 to 18 from the years 2001 to 2023 across various swimming disciplines as our foundation. This analysis process resulted in 16 birth years (ranging from 1990 to 2005), which were examined over an 8-year span across 17 disciplines (including 50m, 100m, and 200m butterfly; back and breaststroke; 200m and 400m individual medley, as well as 50m, 100m, 200m, 400m, and 1500m freestyle). The 800m freestyle was excluded from consideration, as this event has only been part of the Olympics since 2020, thus lacking comparable times throughout the entire observation period. Additionally, the 1500m freestyle was only evaluated from the age of 13, given that the number of competitors in the younger age groups is considerably lower.

This procedure resulted in a dataset comprising a total of 193,920 swimming times from 12,005 athletes. Each accomplishment was distinctly linked to a swimmer according to their name, birth year, and swimming club, and was anonymized using a unique identifier. First, the competition result data (mm:ss) are converted into data with s as the unit; then, the data are normalized to be limited within the interval [0, 1] to ensure the model converges against the effect of outliners. The data normalization procedure is formularized as:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad i = 1, 2, 3, \ldots, n$$

Further Performance metrics were documented based on FINA points, ranking in the annual top 100 list, and position among peers throughout the entire observation period. The predictor variables utilized included:

- Entry_age: age at which the athlete first entered the world's top 100 in the swim discipline.
- Nb_disc_year: Number of different swimming disciplines in the top 100 rankings achieved this year.
- Max_Nb_disciplines: The maximum number of different disciplines within one year.
- dist_cat: distance categories (short 50m, mid 100 and 200m, long 400 and 1500m).
- Nb_dist_cat_year: Number of different distance categories within one year.
- Max_Nb_distance_cat: The maximim number of different distance categories within one year.
- Nb_years_top100: Number of years in the top 100 list, regardless of swimming discipline.
- Nb_years_top100_disc: Number of years in the top 100 list in this discipline.
- prog_year: Percentage improvement over the previous year.
- max_annual_prog: The best improvement rate before the age of 16.
- Max_annual_prog_z_time: The maximum improvement rate for the Z-rate.
- Fina_points_16y:The Fina points scored at the age of 16 in this discipline

The number of observations in each of the 16 events entered over the 8 year analysis period are described in *Table 1*.

To examine the evolution of swimming performance across the years, I employed a mixed or multilevel model (MLM) (Dormehl et al., 2016). This approach facilitates an in-depth analysis of data gathered from the same participants over an extended duration.

Table 1
Cumulative number of performance (between 2001 and 2023) for male swimmers between the age of 11 and 18 years in each event

Number of Performances (years)	50 B	50 F	50 R	50 S	100 B	100 F	100 R	100 S	200 B	200 F	200 L	200 R	200 S	400 F	400 L	1500 F
1	1638	1476	1502	1458	1554	1573	1449	1425	1784	1497	1406	1607	2036	1312	1666	1449
2	798	792	824	744	799	1092	730	736	841	698	696	783	835	712	778	735
3	588	530	495	542	557	606	508	484	529	506	476	504	506	489	519	523
4	400	442	432	421	375	491	382	387	409	383	404	437	384	420	409	445
5	313	335	353	335	335	305	353	307	323	326	330	329	313	304	345	272
6	271	245	272	281	239	223	264	282	251	288	303	245	224	327	277	207
7	202	219	207	225	215	180	228	255	204	229	225	204	211	0	181	0
8	195	210	191	198	220	125	226	228	143	229	228	186	77	0	88	0

The time variable was centered around the initial observation point (11 -12 years), while the progression of performance was represented by a quadratic function that outlines the overall trend within the entire study population. Additionally, this model accommodates individual variations from the general trend, allowing for a thorough depiction of performance evolution. The elevated interclass correlation coefficient (ICC) determined by this model indicates the extent of total variance attributed to differences among individuals, thereby accurately capturing both consistent developmental patterns and individual distinctions. This reinforces the credibility of the analysis and renders the model especially effective for examining long-term advancements in swimming performance. By utilizing the computed quadratic function, I successfully identified the "age of peak performance" for my dataset.

3.1 Missing Data

After refining the data set, a total of 91,734 swimming accomplishments remained from athletes aged 18, which were utilized for subsequent analysis. Missing values in numeric variables were substituted with the median of the corresponding column to prevent distortions stemming from outliers. This method enabled the retrieval of the maximum number of data sets while preserving the integrity of the analysis.

3.2 Data preparation

Once the data set was cleaned, the numeric variables were standardized to ensure they conformed to a uniform scale. In order to accomplish this, the scale() function was employed, which normalized the variables to exhibit a mean of 0 and a standard deviation of 1. The standardized variables comprised:

- FINA points
- maximum annual performance improvement
- number of disciplines
- number of distance categories.

Furthermore, categorical variables, such as swimming disciplines, were transformed into dummy variables through one-hot encoding, making them suitable for analysis.

3.3 Feature Selection

When using ML algorithms for modeling, one ness first to filter out the optimal features to improve the performance of model prediction. If all features are included, it will increase the computational complexity and reduce the model performance. Hence, dimensionality reduction becomes the key to solving the problem. In this study, I used lasso regression as a key method for feature selection and dimensionality reduction. This technique is particularly effective in improving the interpretability and performance of models by imposing a 1-norm constraint on the regression coefficients, which is the sum of their absolute values. This constraint helps to eliminate less important features while retaining the most relevant features, thereby simplifying the model and avoiding overfitting. The application of lasso regression in this study was crucial for filtering characteristics in the prediction models for the swimmers who scored 750 or more Fina points at age 18.

3.4 Machine Learning Model Building and Verification

Three versions of the 750 Fina points prediction model are developed using Logistic Regression, K-nearest neighbor and Random Forest algorithms.

The output of the model is whether the 18 years old athlete can achieve 750 Fina Points or not. I use the fivefold cross-validation method to verify the model's performance. The specific process was splitting the dataset into five groups and assigning them each to an independent folder, four groups used as training data for building the model, and the remaining one used as test data to verify the model's effectiveness. Then, this process was repeated five times, and each of the five verifications was used as the results only once. Then take the average of the five results to get an estimate.

Logistic regression was used to calculate the probability that an athlete would fall into one of the two groups. Multinomial logistic regression is particularly well-suited for predicting multiple categories, as it calculates the probability of an athlete being part of one of these three groups (Denham, 2016). Logistic regression uses the L2 penalty logistic regression

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right)$$

The Random Forest is an ensemble learning algorithm that builds predictions by aggregating the outputs of multiple decision trees. It provides robust predictions by combining the results from several individual trees. Employs the Bootstrap technique to draw n samples from the sample set and creates n classification trees to establish a random forest (Breiman, 2001; Austin et al., 2013). The outcome of the voting process from the classification tree dictates the classification prediction result for the new data as illustrated by the following equation:

$$f(x) = \arg\max_{Y} \sum_{i=1}^{n} \mathbb{I}(h_i(X) = Y)$$

Where hi denotes the fundamental model of an individual classification tree, Y signifies the output variable, and I refers to the indicative function.

The KNN function can be expressed as (Euclidean distance):

$$p_{ij} = \frac{\exp\left(-\left(Lx_i - Lx_j\right)^2\right)}{\sum_{k \neq i} \exp\left(-\left(Lx_i - Lx_k\right)^2\right)}, \quad p_{ij} = 0$$

3.5 Model Evaluation

Evaluation metrics encompass the area beneath the receiver operating characteristic curve (ROC) AUC, along with accuracy, sensitivity, precision, and the balanced F1 score. AUC serves to assess the model's ability to discriminate and its overall performance. An AUC value of 1 indicates the model is flawless; conversely, a value of 0.5 suggests the inadequate performance of a random classifier, meaning that it lacks any discriminative capacity; values ranging from 0.90 to 1 denote excellent performance, 0.80 to 0.90 signifies good performance, 0.70 to 0.80 is considered fair, 0.60 to 0.70 indicates poor performance, and 0.50 to 0.60 represents failure (Peter & Gedeck, 2020).

The correct rate is the proportion of the samples judged correctly by the classifier among all samples.The higher the correct rate, the better the classifier; sensitivity is the proportion of all positive examples judged correctly by the classifier, which measures the classifier's ability to recognize positive examples; accuracy represents the proportion of positive examples judged to be positive by the classifier; the F1 score is the harmonic mean of precision and recall, providing a balanced measure of model performance.; the

maximum of the four indicators is 1, the minimum is 0, and the higher the value, the better the model (Stehman, 1997).

Among the results of judgment, TP=true positive, TN=true negative, FP=false positive, FN=false negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = 2 \times \frac{TP}{2 \times TP + FP + FN}$$

3.6 Class Weighting

To compensate for class imbalance and prevent bias toward the more frequent class, **class weighting** was applied. The weights were assigned inversely proportional to the number of samples in each class, giving more importance to the less represented class. This ensured that the model treated both classes with similar priority, leading to more balanced predictions.

The software R was used for all these methods.

4 Statistical analysis

4.1 Variance and Coefficient of Variation

The consistency of performance within a specific age cohort can be assessed by evaluating the variance of times across various disciplines. For each age category x, the variance of the times for all swimmers within that category is computed. The mean of the times mean_x and the variance of the times Var_x for the age cohort x are determined as follows:

$$\text{mean}_x = \frac{1}{N} \sum_{i=1}^{N} t_{xi}$$

$$\text{Var}_x = \frac{1}{N} \sum_{i=1}^{N} (t_{xi} - \text{mean}_x)^2$$

Where t_{xi} represents the time recorded by swimmer i at age x, and N signifies the total number of swimmers in this age category. The variance Var_x reflects the extent to which the times of swimmers in the age category x deviate from the average.

The coefficient of variation (CV) serves as a standardized indicator of relative variability and is calculated in the following manner:

$$\text{CV}_x = \frac{\text{Std}_x}{\text{mean}_x}$$

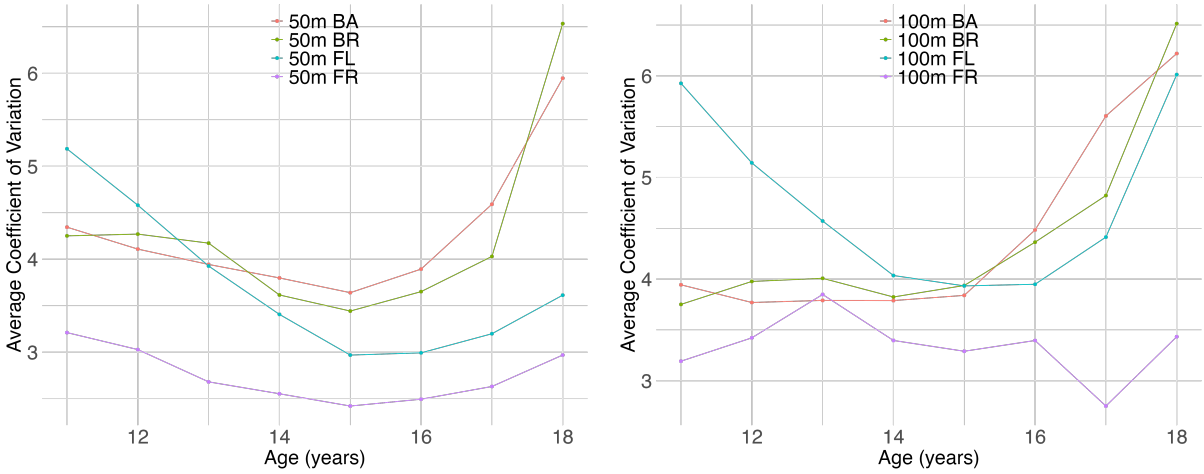
In this context, Std_x denotes the standard deviation of the times within the age category x, while mean_x is the average of the times for this age group. The CV illustrates the magnitude of the relative spread of performance in relation to the mean value.

Essentially, a lower variance signifies a more consistent performance at a given age. The CV illustrates the relative variability of performance within an age group. For

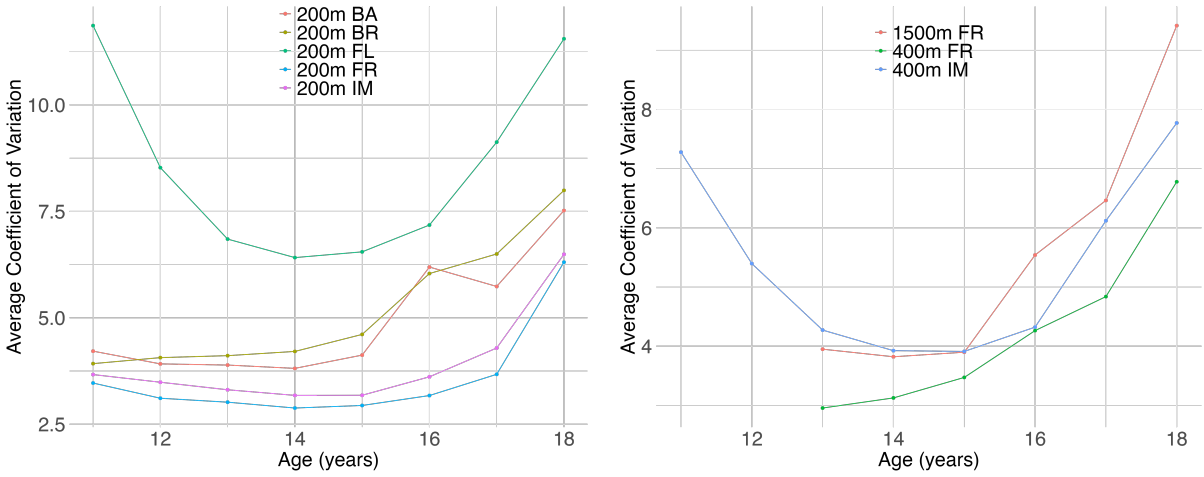
the analysis, the average variance and CV for various swimming events were computed and visually represented.

Figure 1

The CV as a function of age for different stroke over different distances



(a) *50m* **(b)** *100m*



(c) *200m* **(d)** *400+m*

Based on figure 1, we can draw the following conclusions:

1. For all examined swimming distances (50m, 100m, 200m, and 400-1500m), the coefficient of variation (CV) displays a notable increase starting around the age of

15. This indicates that the variation in performance among older age groups is expanding, possibly attributed to differing developmental stages and specializations among swimmers. However, the attrition of more dominant swimmers may also contribute to this trend, resulting in a less pronounced power density within the top 100.

2. In the middle age groups (12-15 years), the CV remains fairly consistent across all distances or even exhibits a slight decrease. This may suggest that a certain level of uniformity in swimming performance is reached within this age group, likely due to a high power density.
3. Particularly in more technically and physically challenging events like the 200m butterfly or 400m freestyle, the CV values are elevated compared to other events, especially within the 11-12 age group. This may imply that swimmers in these events experience larger disparities in their technical performance and their ability to adapt to the demands of the discipline.

4.2 Pearson correlation

The Pearson product-moment coefficient is a measure of the linear correlation between two variables X and Y and takes values between -1 and +1, where +1 represents the strongest positive correlation, 0 is no correlation, and -1 represents the strongest negative correlation. This quantity is widely used as a measure of linear dependence or association between two variables. Here, I use this quantity to study the potential correlation between swimmers' performances at age 18 and at younger ages.

For a given age group, let $X = [x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{18}]$, where x_a is the average swimming time at the age of a. I use the Pearson correlation coefficient to measure the degree of swimming performance correlation between the average times at age x_a (where $a \in [11, 17]$) and x_{18}, defined as

$$\rho_{x_a,x_{18}} = \frac{\text{cov}(x_a, x_{18})}{\sigma_{x_a}\sigma_{x_{18}}}$$

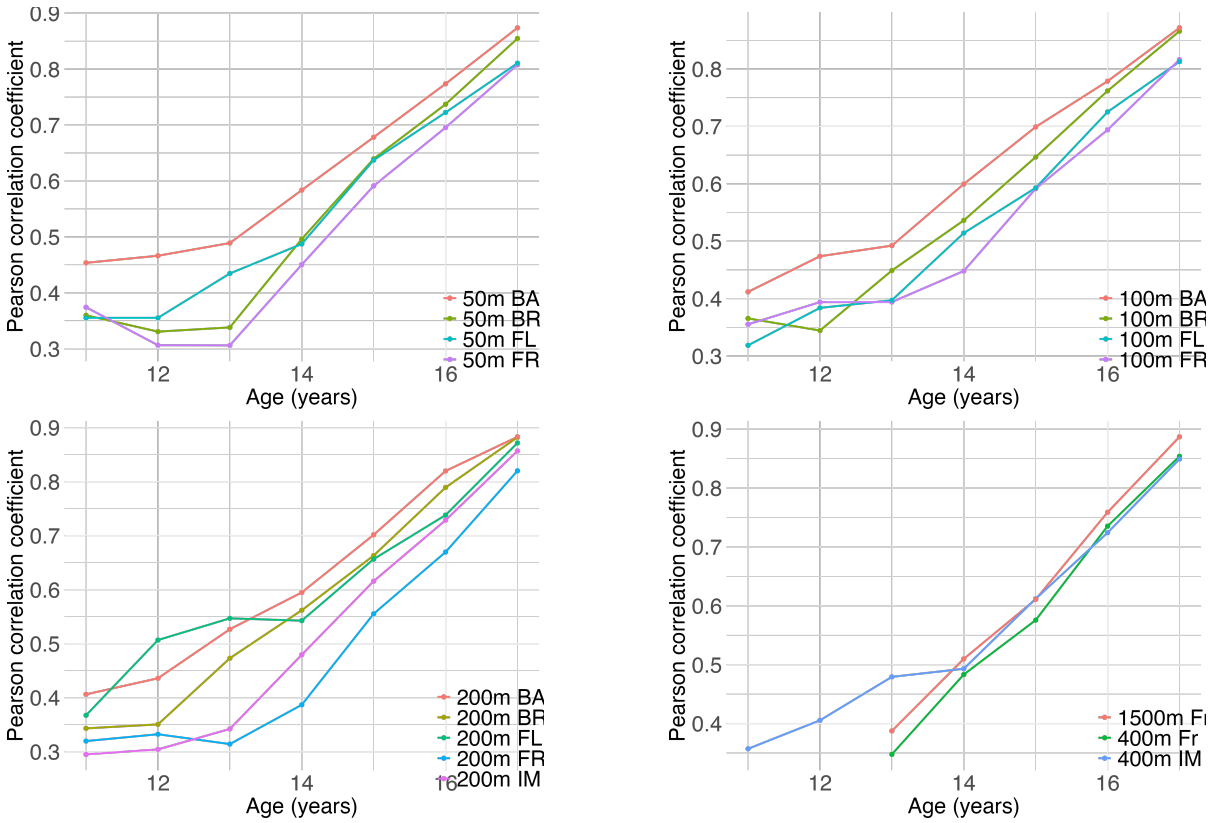
where cov is the covariance and σ_x is the standard deviation of x. Typically, a graded interpretation of the correlation strength (based on Dancey Reidy's 2004 categorization) is as follows:

- 0.0-0.2 = weak or zero correlation
- 0.2-0.4 = modest
- 0.4-0.6 = moderate
- 0.6-0.8 = strong
- 0.8-1.0 = very stron

Figure 2 shows the Pearson correlation between the swimmer's performance at age 18 and those at younger ages. For all events, age 17 showed the strongest correlation with age 18. At younger ages (11 to 13years) the Pearson correlation coefficient is almost steady in the 100m and 200m events, indicating that the developing male body has little to do with performance at young ages. After age 13, for all strokes, the Pearson correlation coefficient steadily increases. Coaches can use these correlations to determine when an athlete's performance reliably predicts future success. A strong correlation after age 14 indicates that tailored training can effectively prepare athletes for peak performance. In summary, Pearson correlation values reveal that swimming performance becomes a more reliable predictor of future success between the ages of 14 and 15. This insight is essential for effectively planning and adapting training to enhance performance outcomes.

Figure 2

Pearson Correlation coefficient between younger ages and age 18.



4.3 Shapiro-Francia Test and Multi-level Model

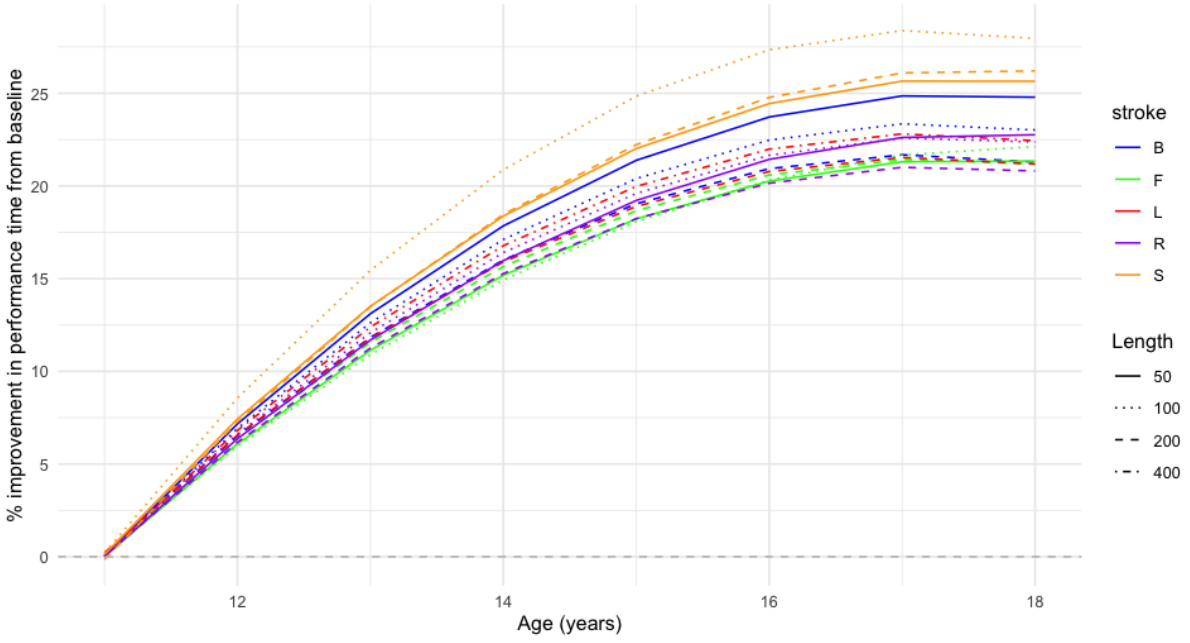
The raw datasets for all performances across the 16 events were evaluated for normality through the Shapiro-Francia test. All event datasets exhibited non-normal distributions. The trajectories of the curves depicting performance progression throughout adolescence were examined utilizing mixed or multi-level modeling (MLM) in R. Time was standardized to zero at the initial observation point (11 years of age), employing an unstructured covariance approach. To ensure thoroughness, the first observation point was also evaluated at ages 12, 13, and 14. The fit of the models pertaining to fixed and random effects was assessed by extracting maximum likelihoods through a hierarchical approach. The ultimate models were represented as quadratic functions for fixed effects (y = ax2 + bx + c). The fixed effects of time illustrated polynomial transformations of the population as they aged, while the random effects indicated individual variations from the mean trajectory of the sample. Inter-class correlations were com-

puted, and R2 values were established to quantify the differences in variability between individuals and within-person effect sizes, respectively.

A noticeable trend can be seen in models. Peak performance are attained in all disciplines when athletes reach the age of 17 to 18. The progression of butterfly swimmers, especially in the 100m event, was distinct from other disciplines. Butterfly swimmers exhibited a significantly greater rate of improvement. Conversely, freestyle swimmers demonstrated the slowest and least amount of progress. Freestyle is primarily practiced during training sessions, which means that the younger swimmers in this case possess superior technical and physical training compared to those in other disciplines.

Figure 3

Quadratic function of the percentage improvement (base age 11 years).

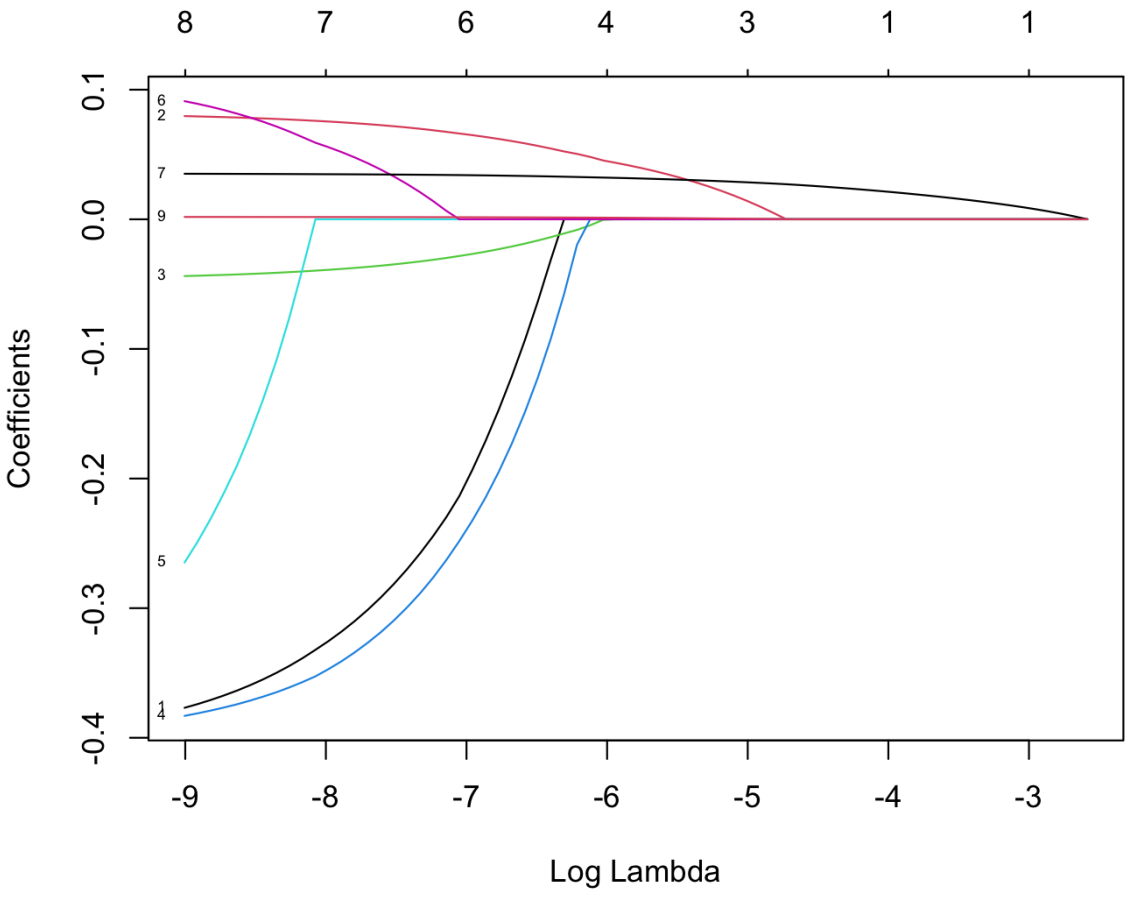


5 Results

5.1 Feature Selection Results of Lasso Regression

Features other than those linked to the 750 FINA points were refined through Lasso regression analysis to identify the optimal features for model development. When λ was set to 0.0001228 (**Figure 4**), the model based on the following features performed best: Early entry age, improvement rate, maximum yearly improvement, years in top 100, maximum number of distance categories, maximum number of different disciplines, FINA points at the age of 16, and z-time ratio.

Figure 4
the path diagram of lasso regression coefficients.



5.2 Performance Prediction Model Results

According to the plotted ROC curve, the AUC values of the instances of the 750 Fina points model for 18 years old athletes established by RF, LR and KNN are 0.97, 0.94, 0.93, respectively. It can be observed that the Random Forest model, with an AUC of 0.97 and an accuracy of 0.93, demonstrates the best overall performance, with the high recall of 0.96 indicating strong detection of relevant cases (less than 750 FINA points), while the F1 score was also solid at 70.3. In comparison, the K-Nearest-Neighbor model also achieves solid results, with an AUC of 0.94 and a high accuracy of 0.98, though its recall (0.69) suggests lower identification of relevant cases. In contrast, the Logistic Regression model shows the weakest performance, with a comparable AUC of 0.94 but significantly lower accuracy (0.02), precision (0.17), and F1-score (0.02), making it less suitable for this classification task. Overall, the random forest showed the best results due to its robust performance and high precision, while the KNN model achieved high accuracy but performed weaker in recognising relevant cases. Logistic regression proved to be unsuitable for this prediction, indicating a strong disparity between the models.

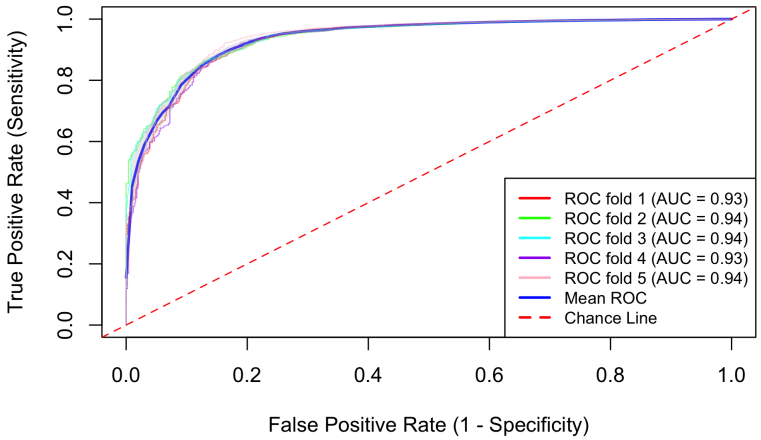
5.3 Predictive Performance

The Logistic Regression model showed an overall accuracy of 96.62 on the test data, with a precision of 72.89 for Group 1 and 97.91 for Group 2. The F1-score was 38,20 for Group 1 and 98.881 for Group 2. These results indicate that the model is significantly better at predicting athletes in Group 2 than in Group 1, which can be attributed to the larger number of athletes with less than 750 FINA points.

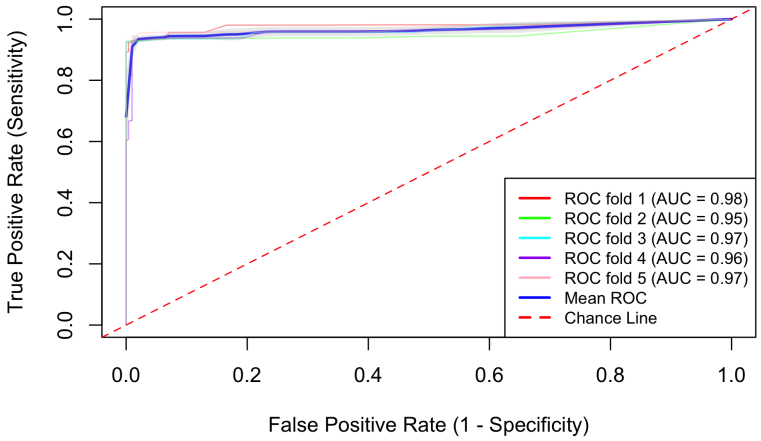
The Random Forest model achieved a higher overall accuracy of 99.84, with a precision of 100 for Group 1 and 99.83 for Group 2. The F1-score was 96.99 for Group 1 and 99.92 for Group 2. This shows that the Random Forest model achieved higher performance in predicting both groups.

Figure 5

Receiver operating characteristic (ROC) curves of different models in the 750 Fina Points prediction. ***(a)*** *LR: logistic regression;* ***(b)*** *RF: random forest;* ***(c)*** *KNN: k-nearest neighbor.*



(a) ***LR***



(b) ***RF***

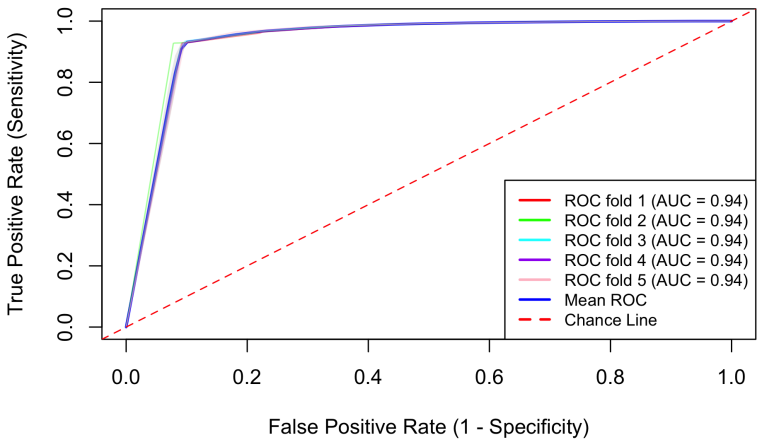


Table 2
Validity evaluation of different prediction models for 750 Fina at the age of 18 years.

ML	Accuracy	Sensitivity	Precision	F1 Score
RF	0.92 ± 0.04	0.96 ± 0.06	0.64 ± 0.03	0.70 ± 0.01
LR	0.02 ± 0.02	0.37 ± 0.05	0.17 ± 0.08	0.02 ± 0.05
KNN	0.98 ± 0.01	0.69 ± 0.01	0.83 ± 0.04	0.74 ± 0.02

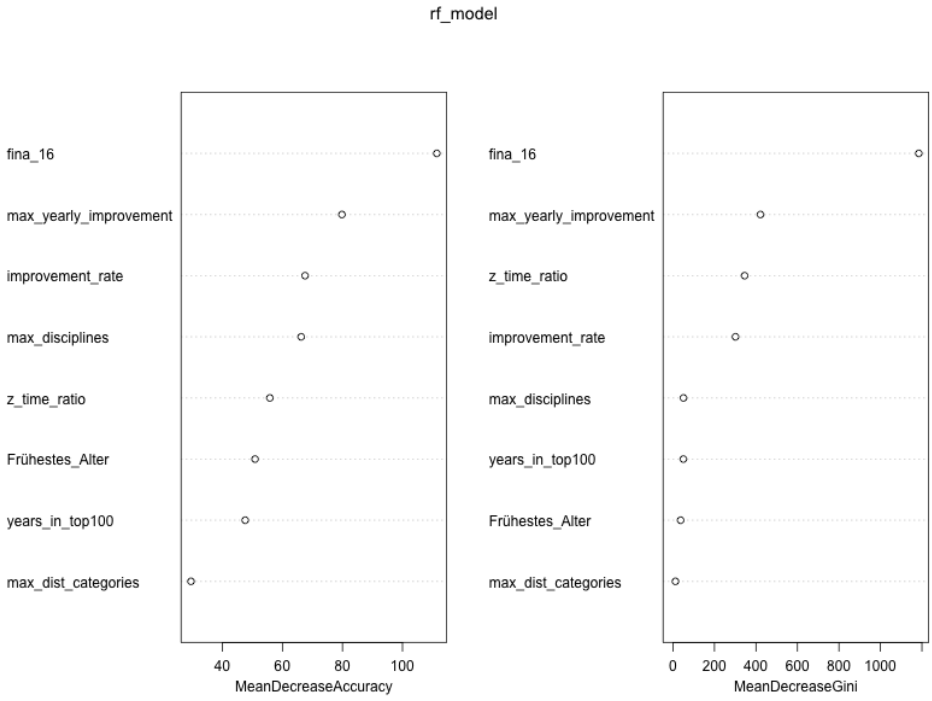
RF, random forest; LR, logistic regression; KNN, K-nearest neighbor.

5.4 Relative Strength of Variables

The feature importance was calculated for the Random Forest model, while the coefficients were analyzed for the Logistic Regression.

Figure 6 shows the relative importance of the variables for the Random Forest model.

Figure 6
Relative importance of the variables (RF).



5.5 Interpretation of the Coefficients

The coefficients can be interpreted by examining the increase or decrease in the odds when the unstandardized variables vary. When a variable X is increased by Δ, the odds of being in Group 1 are multiplied by the factor $e^{(B\Delta/\text{std})}$, where B and std represent the coefficient and standard deviation of X, respectively. The following results are obtained from **Table 3** :

- Early_Age (average 14, std 2): Lowering the age of entry into the top 100 by 2 years increases the odds of being in group 1 by a factor of 2.51 (95% CI: 1.36—4.55).
- improvement_rate (avg 5, std 2): A 2% rise in improvement reduces the odds of being in group 1 by a factor of 0.71 (95% CI: 0.66—0.76).
- max_yearly_improvement (avg 10, std 3): A 3% increase in the maximum yearly improvement raises the odds of being in group 1 by a factor of 1.14 (95% CI: 1.03—1.27).
- years_in_top100 (avg 4, std 2): Spending an additional year in the top 100 boosts the odds of being in group 1 by a factor of 1.65 (95% CI: 1.19—2.30).
- max_dist_categories (avg 2, std 1): Increasing the maximum distance categories by 1 enhances the odds of being in group 1 by a factor of 1.15 (95% CI: 0.78—1.70).
- max_disciplines (avg 3, std 1): A rise in the maximum disciplines yields a slight negative effect with a factor of 0.90 (95% CI: 0.73—1.08), though this is not statistically significant.
- fina_16 (avg 750, std 50): An increase of 50 points in FINA scores at age 16 reduces the odds of being in group 1 by a factor of 0.07 (95% CI: 0.06—0.08), indicating a significant negative effect.
- max_z_time_rate (avg 0.8, std 0.1): This variable demonstrated considerable uncertainty with a factor of 59.26 (95% CI: 0.00—9.72e+09).

- z_time_ratio (avg 1, std 0.2): A decrease of 0.2 in the Z-time ratio reduces the odds of being in group 1 by a factor of 0.92 (95% CI: 0.89—0.96).

Table 3
The ten of most relevant variables (among the initial 12), in standardized form obtained through exhaustive feature selection and five-fold cross-validation on the training set.

Variable	**B**	**SD**	**Z-Value**	**p-value**	**95% CI (B)**	**e^B**	**95% CI (e^B)**
Intercept	6.08	0.27	22.52	<0.001	(5.55 : 6.61)	438.87	(257.15 : 744.97)
Early_Age	0.92	0.31	2.97	<0.01	(0.31 : 1.52)	2.51	(1.36 : 4.55)
improvement_rate	-0.34	0.04	-7.81	<0.001	(-0.42 : -0.27)	0.71	(0.66 : 0.76)
max_yearly_improvement	0.13	0.05	2.47	<0.05	(0.03 : 0.24)	1.14	(1.03 : 1.27)
years_in_top100	0.50	0.17	2.96	<0.01	(0.17 : 0.83)	1.65	(1.19 : 2.30)
max_dist_categories	0.14	0.20	0.70	0.08	(-0.25 : 0.53)	1.15	(0.78 : 1.70)
max_disciplines	-0.11	0.10	-1.15	0.12	(-0.31 : 0.08)	0.90	(0.73 : 1.08)
fina_16	-2.63	0.06	-41.25	<0.001	(-2.74 : -2.52)	0.07	(0.06 : 0.08)
max_z_time_rate	4.09	9.51	0.43	0.14	(-14.54 : 22.72)	59.26	(0.00 : 9.72e+09)
z_time_ratio	-0.08	0.02	-3.58	<0.05	(-0.12 : -0.04)	0.92	(0.89 : 0.96)

6 Discussion

Machine Learning Models

This study investigated three distinct machine learning methods utilizing real competition data from the top 100 rankings of male adolescent swimmers. The findings indicate that swimming performance can indeed be predicted using machine learning algorithms. Notably, the random forest model (RF) demonstrated exceptional capability, achieving an AUC of 0.97 and an accuracy of 92.97. This impressive predictive accuracy emphasizes the model's potential to furnish athletes and coaches with crucial insights early on, facilitating strategic decisions and the development of tailored training programs.

The strength of the RF model is further evidenced by its high recognition rate of pertinent cases (recall 96.24), suggesting that it is especially adept at pinpointing swimmers who possess the potential to attain over 750 FINA points. This is vital in practical terms, as such predictions enable coaches to monitor the performance trends of young athletes more closely and customize training to optimize long-term success.

Another benefit of the RF approach is its capacity to assess the significance of individual variables, offering valuable insights into the factors that most significantly impact athletic achievement. This capability not only allows for more effective training oversight but also facilitates the early identification of swimmers' strengths and weaknesses. When compared to alternative models (such as Logistic Regression and KNN), Random Forest exhibits considerable superiority and delivers more dependable predictions that can greatly assist both short-term and long-term training strategies.

The K-nearest neighbor model (KNN) demonstrated commendable performance with impressive predictive accuracy.Nevertheless, in comparison to alternative models,

it exhibited a reduced capacity to identify all pertinent cases, especially swimmers who earn over 750 FINA points. This implies that although KNN performs admirably in overall classification, it is less adept at comprehensively recognizing potential elite athletes.

In practical scenarios, this suggests that the KNN model could be especially beneficial when exceptionally high precision and accuracy are essential, such as in making definitive decisions regarding athletes who are already well-established.However, the diminished recall rate suggests that it is not as effective for spotting young talents that are still in their developmental stages. In training contexts where the goal is to uncover as many prospective talents as possible, the KNN model may thus prove less efficient compared to other models like Random Forest. Nonetheless, due to its straightforward implementation and notable precision, KNN presents a valuable strategy for classification in specific circumstances.

Logistic regression exhibited the least effective performance in this research when compared to the other models. With a notably lower accuracy of 2.4, alongside low precision at 16.9 and an F1 score of 23.7, this model struggled to sufficiently predict swimming performance. A primary factor contributing to this issue is the linear nature of the model, which finds it challenging to capture the intricate and nonlinear relationships among the variables that are characteristic of athletic performance.

The parameters examined in the study, including the annual maximum improvement (maxyearlyimprovement) and the number of years ranked in the top 100 (yearsintop100), frequently do not follow a linear correlation. A linear model like logistic regression fails to effectively portray such interactions. Specifically, factors such as the improvement rate or the diversity of disciplines exhibit a variability that necessitates nonlinear modeling.

Moreover, the imbalanced distribution of data posed a significant challenge. There was a disproportionate number of swimmers with less than 750 FINA points, resulting in the logistic regression establishing an inaccurate decision threshold. Even after applying class weighting, the model could not match the distinguishing ability of more sophisticated models like Random Forest.

While certain parameters, such as FINA points at the age of 16 years (fina16), were managed relatively well by logistic regression due to their clear linear connection to the target variable, other variables like the Z-time ratio were more challenging to represent. Logistic regression struggled to account for nonlinear influences, which resulted in subpar overall predictive performance.

Another crucial aspect is the interplay between the variables. Elements such as the earliest age of entry into the top 100 and annual improvement may involve intricate interactions that logistic regression was unable to capture. Consequently, the model remained constrained in its capacity to accurately forecast swimming performance.

In summary, it seems that logistic regression is not well-suited for predicting swimming performance, where numerous factors exhibit complex interrelationships. Algorithms that are more adept at detecting such nonlinear patterns, such as Random Forest, yield considerably more dependable results in this context.

Predictors

A particularly significant predictor in the random forest model was the FINA points score achieved at the age of 16 (fina16). This metric demonstrates a robust correlation with performance at 18 years of age. From a critical perspective, this is not surprising, as the FINA points at 16 are relatively similar to those at 18. Athletes who attain a high score at 16 logically possess better prospects for success at 18 years of age due to their enhanced performance and maturity. Conversely, it is statistically

improbable that an athlete who struggles at 16 will exhibit a significant performance surge within two years. This indicates that this factor facilitates short-term forecasting rather than capturing long-term trends or potential development. In subsequent studies, it would be essential to utilize the FINA points of younger athletes for generating long-term predictions.

Another crucial element was the annual performance improvement (improvementrate) and the maximum annual improvement (maxyearlyimprovement). These factors carry greater significance as they illustrate the rate of progress over multiple years. Athletes who demonstrate consistent improvement are thereby indicating their capacity to effectively respond to training and competition demands, which may serve as a more reliable indicator of long-term success.

The age at which athletes enter the top 100 (entryage) was also a pertinent factor. Swimmers who broke into the top 100 at a young age were more likely to achieve over 750 FINA points by 18 years of age. This implies that early athletic successes and a swift establishment at the national level are crucial for long-term achievement.

In contrast, factors like the number of different distance categories (maxdistcategories) or the years spent in the top 100 (yearsintop100) were deemed less significant. This illustrates that it is feasible to attain exceptional performance even in the absence of extensive diversification in disciplines or prolonged placements within the top 100.

7 Strengths, limitation and future research

The random forest models demonstrated a remarkable level of robustness, as they identified nonlinear relationships among the variables, thereby facilitating an accurate prediction of swimming performance. Moreover, significant factors like FINA points at the age of 16 and annual improvement yield critical insights for training development. The restriction of data to male athletes constrains the applicability of the findings. Additionally, utilizing FINA points at 16 years old allows for primarily short-term forecasts. The choice of parameters could be refined further, and incorporating additional variables might lead to the creation of new and potentially superior predictive models.

Future investigations should encompass female athletes and younger age groups to deliver more extensive predictions.Furthermore, integrating training and physical performance metrics, such as body size, strength, or endurance, could enhance the models and address more dimensions of competitive sports training for young talent.

8 Conclusion

This study has demonstrated that the random forest model provides exceptional predictive accuracy, with key factors such as FINA points and the rate of improvement being crucial for success compare with KNN and LR. Young male swimmers who achieve a very high score and maintain a consistent improvement rate are likely to attain a high score at the age of 18. The findings offer valuable insights for the formulation of training strategies and pave the way for new avenues in predicting swimming performance.

Bibliography

Abbott, S., Hogan, C., Castiglioni, M. T., Yamauchi, G., Mitchell, L. J., Salter, J., Romann, M., & Cobley, S. (2021). Maturity-related developmental inequalities in age-group swimming: The testing of 'mat-caps' for their removal. *24*(4), 397–404. https://www.sciencedirect.com/science/article/pii/S1440244020307829

Allen, S. V., & Hopkins, W. G. (2015). Age of peak competitive performance of elite athletes: A systematic review. *Springer Science+Business Media*, *45*(10), 1431–1441. https://doi.org/10.1007/s40279-015-0354-3

Allen, S. V., Vandenbogaerde, T. J., & Hopkins, W. G. (2014). Career performance trajectories of olympic swimmers: Benchmarks for talent development. *Taylor & Francis*, *14*(7), 643–651. https://doi.org/10.1080/17461391.2014.893020

Allen, S. V., Vandenbogaerde, T. J., Pyne, D. B., & Hopkins, W. G. (2015). Predicting a nation's olympic-qualifying swimmers. *10*(4), 431–435. https://journals.humankinetics.com/view/journals/ijspp/10/4/article-p431.xml

Alshdokhi, K. A., Petersen, C., & Clarke, C. (2020). Improvement and variability of adolescent backstroke swimming performance by age. *Frontiers Media*, *2*. https://doi.org/10.3389/fspor.2020.00046

Berthelot, G., Johnson, S., Noirez, P., Antero, J., Marck, A., Desgorces, F.-D., & Pifferi, F. (2019).

Berthelot, G., Len, S., Hellard, P., Tafflet, M., Guillaume, M., Vollmer, J.-C., Gager, B., Quinquis, L., Marc, A., & Toussaint, J.-F. (2011). Exponential growth combined with exponential decline explains lifetime performance evolution in individual and human species. *Springer Science+Business Media*, *34*(4), 1001–1009. https://doi.org/10.1007/s11357-011-9274-9

Boccia, G., Moisè, P., Franceschi, A., Trova, F., Panero, D., Torre, A. L., Rainoldi, A., Schena, F., & Cardinale, M. (2017). Career performance trajectories in track and

field jumping events from youth to senior success: The importance of learning and development. *12*(1), 1–15. https://doi.org/10.1371/journal.pone.0170744

Born, D.-P., Lomax, I., Rüeger, E., & Romann, M. (2022). Normative data and percentile curves for long-term athlete development in swimming. *25*(3), 266–271. https://www.sciencedirect.com/science/article/pii/S1440244021004540

Born, D.-P., Stöggl, T., Lorentzen, J., Romann, M., & Björklund, G. (2023). Predicting future stars: Probability and performance corridors for elite swimmers. *Elsevier BV*. https://doi.org/10.1016/j.jsams.2023.10.017

Bügner, J. (2005). Nichtlineare methoden in der trainingswissenschaftlichen diagnostik - mit untersuchungen aus dem schwimmsport.

Cobley, S., Abbott, S., Dogramaci, S., Kable, A., Salter, J., Hintermann, M., & Romann, M. (2018). Transient relative age effects across annual age groups in national level australian swimming. *21*(8), 839–845. https://www.sciencedirect.com/science/article/pii/S1440244017318662

Costa, M. J., Bragada, J. A., Marinho, D. A., de Ribeiro dos Reis, V. M. M., Silva, A. J., & Barbosa, T. M. (2014). Longitudinal assessment of swimming performance in the 200-m freestyle event. https://doi.org/https://doi.org/10.2174/1875399x010030100092

Costa, M. J., Bragada, J. A., Marinho, D. A., Lopes, V. P., Silva, A. J., & Barbosa, T. M. (2013). Longitudinal study in male swimmers: A hierachical modeling of energetics and biomechanical contributions for performance. https://pubmed.ncbi.nlm.nih.gov/24421719

Costa, M. J., Marinho, D. A., Bragada, J. A., Silva, A., & Barbosa, T. M. (2011). Stability of elite freestyle performance from childhood to adulthood. https://doi.org/https://doi.org/10.1080/02640414.2011.587196

de Jesus, K., Ayala, H. V. H., de Jesus, K., dos Santos Coelho, L., Medeiros, A. I. A., Abraldes, J. A., Vaz, M., Fernandes, R. J., & Vilas-Boas, J. P. (2018). Modelling

and predicting backstroke start performance using non-linear and linear models. *De Gruyter Open*, *61*(1), 29–38. https://doi.org/10.1515/hukin-2017-0133

Denham, B. E. (2016). Multinomial logistic regression. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119407201.ch7

Dormehl, S. J., Robertson, S., & Williams, C. A. (2016). Modelling the progression of male swimmers' performances through adolescence. *Multidisciplinary Digital Publishing Institute*, *4*(1), 2–2. https://doi.org/10.3390/sports4010002

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2001). Prognose der olympischen wettkampfleistung im schwimmen. *Leistungssport*, *31 (3)*, 20–23. https://eref.uni-bayreuth.de/id/eprint/3335

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2006). Modellierung von wettkampfleistung im schwimmen bei den olympischen spielen 2000 und 2004 mittels neuronaler netze. https://eref.uni-bayreuth.de/3115/

Foss, J. L., Sinex, J. A., & Chapman, R. F. (2019). Career performance progressions of junior and senior elite track and field athletes. *Springer Science+Business Media*, *1*(2), 168–175. https://doi.org/10.1007/s42978-019-0013-8

Gulbin, J. P., Croser, M. J., Morley, E. J., & Weissensteiner, J. R. (2013). An integrated framework for the optimisation of sport and athlete development: A practitioner approach. *31*(12), 1319–1331. https://doi.org/10.1080/02640414.2013.781661

Haar, B. (2011). Analyse und prognose von trainingswirkungen: Multivariate zeitreihenanalyse mit künstlichen neuronalen netzen. https://doi.org/10.18419/opus-5549

Haugen, T., Solberg, P., Foster, C., Morán-Navarro, R., Breitschädel, F., & Hopkins, W. G. (2018). Peak age and performance progression in world-class track-and-field athletes. *13*(9), 1122–1129. https://journals.humankinetics.com/view/journals/ijspp/13/9/article-p1122.xml

Hohmann, A., Edelmann, J., & Henneberg, B. (2000). A nonlinear approach to the analysis and modeling of training and adaptation in swimming. *1*(1). https://ojs.ub.uni-konstanz.de/cpa/article/download/2529/2376

Kozieł, S. M., & Malina, R. M. (2017). Modified maturity offset prediction equations: Validation in independent longitudinal samples of boys and girls.

Malcata, R. M., Hopkins, W. G., & Pearson, S. (2014). Tracking career performance of successful triathletes. *Lippincott Williams & Wilkins*, *46*(6), 1227–1234. https://doi.org/10.1249/mss.0000000000000221

Monteiro, D., Marinho, D. A., Moutão, J., Vitorino, A., Antunes, R., & Cid, L. (2017). Reasons for dropout in swimmers, differences between gender and age and intentions to return to competition. *58*(1-2), 180–192. https://doi.org/10.23736/s0022-4707.17.06867-0

Mujika, I., Pyne, D. B., Wu, P., Ng, K., Crowley, E., & Powell, C. (2023). Next-generation models for predicting winning times in elite swimming events: Updated predictions for the paris 2024 olympic games. *Human Kinetics*, *18*(11), 1269–1274. https://doi.org/10.1123/ijspp.2023-0174

Peter, A. B., & Gedeck, P. (2020). Practical statistics for data scientists 50+ essential concepts using r and python.

Pyne, D. B., Trewin, C., & Hopkins, W. G. (2004). Progression and variability of competitive performance of olympic swimmers. https://doi.org/10.1080/02640410310001655822

Sokolovas, G. (2006). Analysis of usa swimming's all-time top 100 times.

Staub, I., Zinner, C., Bieder, A., & Vogt, T. (2020). Within-sport specialisation and entry age as predictors of success among age group swimmers. *Taylor & Francis*, *20*(9), 1160–1167. https://doi.org/10.1080/17461391.2019.1702107

Staub, I., Zinner, C., Stallman, R. K., & Vogt, T. (2019). The consistency of performance among age group swimmers over 8 consecutive years. *Springer Nature*, *50*(1), 123–129. https://doi.org/10.1007/s12662-019-00628-8

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *62*(1), 77–89. https://www.sciencedirect.com/science/article/pii/S0034425797000837

Svendsen, I. S., Tønnesen, E., Tjelta, L. I., & Dickstein, K. (2018). Training, performance, and physiological predictors of a successful elite senior career in junior competitive road cyclists. *13*(10), 1287–1292. https://doi.org/10.1123/ijspp.2017-0824

Tchamkerten, A., Chaudron, P., Girard, N., Tchamkerten, A., Tchamkerten, A., & Tchamkerten, A. (2024). Career factors related to winning olympic medals in swimming. *19*(6), 1–15. https://doi.org/10.1371/journal.pone.0304444

Vantorre, J., Chollet, D., & Seifert, L. (2014). Biomechanical analysis of the swim-start: A review. *National Institutes of Health*. https://pubmed.ncbi.nlm.nih.gov/24790473

Walther, J., Mulder, R. C., Noordhof, D. A., Haugen, T., & Sandbakk, Ø. (2022). Peak age and relative performance progression in international cross-country skiers. *17*(1), 31–36. https://journals.humankinetics.com/view/journals/ijspp/17/1/article-p31.xml

Wu, P., Babaei, T., O'Shea, M., Mengersen, K., Drovandi, C., McGibbon, K. E., Pyne, D. B., Mitchell, L. J. G., & Osborne, M. (2021). Predicting performance in 4 x 200-m freestyle swimming relay events. *Public Library of Science*, *16*(7), e0254538–e0254538. https://doi.org/10.1371/journal.pone.0254538

Xie, J., Xu, J., Nie, C., & Nie, Q. (2016). Machine learning of swimming data via wisdom of crowd and regression analysis. *Arizona State University*, *13*(6), 9–9. https://doi.org/10.3934/mbe.2017031

Yustres, I., del Cerro, J. S., González-Mohíno, F., Peyrebrune, M. C., & González-Ravé, J. M. (2019). Comparing the pathway to success in european countries competing in the swimming world championships. *Frontiers Media*, *10*. https://doi.org/10.3389/fpsyg.2019.01437

Yustres, I., del Cerro, J. S., González-Mohíno, F., Peyrebrune, M. C., & González-Ravé, J. M. (2020). Analysis of world championship swimmers using a performance progression model. *Frontiers Media*, *10*. https://doi.org/10.3389/fpsyg.2019.03078

Declaration of Academic Integrity

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here.

This paper was not previously presented to another examination board and has not been published.

Munich, 30.09.2024 Felix-Daniel Bongarz