Is it possible to predict a marathon time based on machine learning?

Afbeelding met persoon, kleding, schoeisel, buitenshuis

Automatisch gegenereerde beschrijving

**Titel Report:** Is it possible to predict a marathon time based on machine learning?

**Authors contact details:** Martijn van der Wielen

[m.vanderwielen2@student.han.nl](mailto:m.vanderwielen2@student.han.nl)

Mariëlle Schouten (1659397)

[m.schouten7@student.han.nl](mailto:m.schouten7@student.han.nl)

**Education:** Minor DDD3

**Course:** Datamining

**Teacher:** Witek ten Hoven

**Submission date:** March 29 2024

**Opportunity:** 1

Inhoud

[1. Project Objective 3](#_Toc158684937)

[1. Business Understanding 4](#_Toc1588307919)

[1.1 Research Question 4](#_Toc436129987)

[1.2. Literature Review 4](#_Toc2002242109)

[1.2.1. The first article: Prediction of Marathon Performance Using Artificial Intelligence 5](#_Toc2032314107)

[1.2.2. The second article: Predicting Race Time in male amateur marathon runners 6](#_Toc1501140644)

[1.2.3. The third article: Application of K-Nearest Neighbors Variants for Disease Prediction 8](#_Toc860597077)

[3. Data understanding 12](#_Toc68377204)

[4. Data Preparation 14](#_Toc292969607)

[5. Model Development and Training 16](#_Toc1623563844)

[5.1. Model Refinement Strategies 17](#_Toc307212346)

[5.2.Training the model and playing with the variables 18](#_Toc2009299849)

[5.3. Development and Evaluation of the Neural Network Regression Model 19](#_Toc1648852320)

[6. Evaluation and Iteration 20](#_Toc1304798977)

[6.1. Similarities 21](#_Toc1977022754)

[6.2. Assessment of Relevant Variables from Theoretical Framework 22](#_Toc1533515750)

[6.4. Naive Bayes 24](#_Toc2059548477)

[6.4.1. Naive Bayes Theory 24](#_Toc1132174749)

[6.4.2. Limitations of Naive Bayes in Marathon Time Prediction 25](#_Toc640174285)

[7. Deployment 26](#_Toc470463189)

# 1. Project Objective

In this research, we seek to analyze the factors that influence marathon finishing times. Our ambition is to create a predictive model where we build a tool that uses the complexity of runners' profiles - including age, gender and historical race performance - to predict marathon results. This research aims to provide athletes and coaches with data-driven insights to improve marathon performance.

Based on the structured methodology of the CRISP-DM framework, our research unfolds in two phases, each dedicated to exploring a different non-linear modelling technique: the K-Nearest Neighbours (KNN) algorithm and Neural Network Regression via the TensorFlow Keras API. These phases represent the possibility of predictive analytics in sport, in line with the course objective of mastering essential data mining skills, particularly in non-linear modelling.

Phase 1: Getting better results with KNN

We started with the KNN algorithm, focusing on the careful selection and standardization of important features to predict marathon finish times. Through a step-by-step process of preparing our data, training the model and making continuous improvements - including removing data that didn't fit well and adding half marathon times - we significantly improved the accuracy of our model. We were able to reduce the average error in our predictions from 12 minutes to 7.5 minutes. This part of the project really showed us how important it is to keep tweaking the model and making sure we're using quality data to get better at predicting marathon times.

Phase 2: Going deeper with neural network regression

After our initial success with KNN, we decided to try something more complex by moving to neural networks, using a multi-layer sequential model setup in the TensorFlow Keras environment. This allowed us to explore deeper into the data, looking for patterns through a model that's built up in layers. We used the Adam optimizer and focused on reducing the mean squared error (MSE) as we went through over 100 rounds of tuning and improving our model. Our goal was to get a really good handle on what affects marathon finish times.

By tackling the challenges of non-linear modelling and following a process of data analysis and improving our models, this project not only achieved our educational goals of becoming better at data mining, but also provided us with valuable insights that could help marathon runners and coaches improve.

# Business Understanding

In the rapidly evolving landscape of sports analytics, the application of data science and machine learning (ML) techniques has become a cornerstone for enhancing athletic performance, strategizing training regimes, and predicting future outcomes. The domain of marathon running, a discipline that combines endurance, strategy, and physical prowess, presents a unique opportunity to leverage these technologies for performance prediction. The ability to accurately forecast marathon times based on a variety of predictors not only aids athletes and coaches in their preparation but also enriches the analytical tools available in sports science. This chapter delves into the business understanding phase of the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, setting the stage for a focused investigation into the predictive capabilities of Artificial Neural Networks (ANN) and K-Nearest Neighbors (KNN) in the context of marathon performance. By exploring the intricacies of marathon running, identifying key performance indicators, and understanding the potential impact of accurate predictions, we aim to bridge the gap between theoretical data science applications and practical athletic performance enhancement.

## 1.1 Research Question

The centerpiece of our research question is designed to find out the complexities of predicting marathon performance through machine learning. The question,

*"How can machine learning techniques, specifically Artificial Neural Networks (ANN) and K-Nearest Neighbors (KNN), be effectively applied to predict marathon performances based on available physiological, demographic, and performance-related data, and how does the predictive accuracy of these techniques compare to actual marathon outcomes?"*

serves as a guiding line for our research. This inquiry not only underscores the technical endeavor of applying ANN and KNN algorithms but also highlights the practical significance of such predictions of marathon running. Through this research question, we aim to uncover the potential of machine learning in sports analytics, offering insights that could transform training strategies and performance predictions in the athletic community.

With the research question firmly established, the next step involves conducting a thorough literature review to anchor our study within the existing body of knowledge. This review will explore previous applications of machine learning in sports performance prediction, identify gaps in the current research landscape, and examine the variables commonly associated with marathon success. By synthesizing these findings, we will lay a solid foundation for our investigation, ensuring that our approach is both innovative and grounded in scientific rigor.

## Literature Review

A literature review will be conducted to gather existing knowledge and research on marathon performance, influencing factors, and prediction models. This includes academic articles, case studies, and existing statistical analyses.

### The first article: Prediction of Marathon Performance Using Artificial Intelligence

Introduction

The marathon, an athletic endurance event of 42.195 km, has evolved since the first modern Olympic Games in 1896 and the first "urban tour" marathon in New York City in 1976 into a global social phenomenon. Interest in these events has grown, along with race times for top runners. This has led to a diversity of goals among participants, ranging from simply finishing the race to breaking personal, national, or world records, often driven by personal or economic reasons.

Key Findings

A recent study by Lucie Lerebourg et al. (2023) explored the application of artificial intelligence techniques, specifically Artificial Neural Networks (ANN) and k-Nearest Neighbors (kNN), for predicting marathon performances. This study is notable for being, to the best of our knowledge, the first to validate these two AI techniques within the discipline of marathon running and to directly compare the accuracy and precision of the predicted performances (Lerebourg et al., 2022).

The study analyzed official French rankings from 10 km road races and marathons in 2019, with a dataset of 820 athletes. The input variables used for both kNN and ANN were the same: 10 km race time, Body Mass Index (BMI), age, and gender, to solve the linear regression problem of estimating marathon race time.

Findings and Conclusions

The study confirmed the validity of both algorithms for predicting marathon performance, with kNN performing better than ANN, having a mean absolute error of 2.4% versus 5.6% for ANN. This suggests that predictions from these artificial intelligence methods could be utilized in training programs and competitions (Lerebourg et al., 2022).

**Discussion and Practical Applications**

The study highlights the potential for using ANN and kNN in predicting marathon performance and provides insight into how these methods can contribute to optimizing training programs and competitive strategies. While both methods were found to be valid and accurate, the higher accuracy of kNN suggests that this model may be superior for performance prediction in marathon running.

Conclusion

The integration of artificial intelligence into sports science, specifically in the prediction of marathon performances, offers new avenues for athletes and coaches to optimize training and competition strategies. The findings of this study support the use of kNN as a preferred method for accurate prediction of marathon performance, which could provide significant benefits for the preparation and strategy determination of long-distance runners.

The findings from the study by Lerebourg et al. (2023) have several implications for our project, which aims to predict marathon performance using artificial intelligence:

1. **Validation of AI Techniques:** The study demonstrates that both Artificial Neural Networks (ANN) and k-nearest Neighbors (kNN) are valid techniques for predicting marathon performance. This confirmation allows us to confidently select either of these AI methods for our project.
2. **Superior Performance of kNN:** Since kNN showed better accuracy than ANN in the study, with a mean absolute error of 2.4% versus 5.6% for ANN, it suggests that kNN might be a more suitable choice for our project if we aim for higher prediction accuracy.
3. **Input Variables for Prediction:** The successful use of specific input variables (10 km race time, BMI, age, and sex) to predict marathon times in the study provides a guideline for selecting the features that could be most relevant and impactful for our model.
4. **Practical Application in Training Programs:** The study highlights the potential of using AI predictions in training programs and competitions. It means that the AI model we develop could be used by athletes and coaches to tailor training sessions, set realistic goals, and optimize race strategies based on predicted marathon performance.
5. **Future Directions:** The research by Lerebourg et al. opens the door for further exploration of other variables that might influence marathon performance, such as physiological, psychological, and environmental factors. This suggests that our project could also consider incorporating a broader range of inputs for even more precise performance predictions.

### The second article: Predicting Race Time in male amateur marathon runners

In summary, the study provides a strong foundation and direction for our project by confirming the effectiveness of using AI for marathon performance prediction, recommending kNN for higher accuracy, and suggesting a set of input variables that have proven to be significant predictors of performance. This enables us to proceed with developing an AI model that not only aims to predict marathon times accurately but also has practical implications for training and competition preparation.

The study "Predictors of Half-Marathon Performance in Male Recreational Athletes" explored the relationship between half-marathon race time and various factors such as training, anthropometry, and physiological characteristics. The goal was to develop a predictive formula for estimating the race time of male recreational runners. This was achieved by analyzing data from 134 male recreational runners who had participated in physical fitness tests, providing insights into their training habits and personal records (Salinero et al., 2017).

Key findings included:

* Performance groups differed significantly in their half-marathon race time, training days, training distance, age, weight (BMI), body fat percentage (BF), and maximal oxygen uptake (VO2max), with faster groups scoring better than slower ones.
* Half-marathon race time was found to correlate with physiological, anthropometric, and training characteristics, indicating that faster runners performed better on these measures.
* The study developed a prediction formula for race time using BMI, VO2max, and weekly training distance as predictors. This formula was validated and showed no bias in predicting race times within the control group.

The study emphasizes the importance of optimized body weight through appropriate exercise and nutrition interventions for recreational runners' performance. It demonstrates that half-marathon performance can be predicted using a combination of anthropometric, physiological, and training characteristics, providing a practical tool for recreational runners and the professionals working with them (Salinero et al., 2017).

Regarding the techniques used for analyzing the datasets, the study primarily focused on statistical methods such as correlation analysis and regression models to develop the prediction formula. The methods mentioned specifically for prediction were stepwise linear regression analysis, rather than machine learning techniques like k-Nearest Neighbors (kNN) or Naive Bayes. These regression analyses were used to identify significant predictors of half-marathon race time and to create a formula that could accurately estimate performance based on those predictors.

This approach underlines the potential of statistical analysis in sports science, particularly in predicting athletic performance based on measurable characteristics. While kNN or Naive Bayes were not used in this study, the methodology applied provides valuable insights into how various factors contribute to running performance and how they can be used to guide training and preparation for half-marathons (Salinero et al., 2017).

The insights from the study "Predictors of Half-Marathon Performance in Male Recreational Athletes" provide a clear direction on the importance of combining training data, physiological measurements, and anthropometric characteristics to develop a robust prediction model. Here's what it means for the project to obtain our main goal:

1. **Data Collection Focus**: We should gather data not only on the runners' race times but also on their training routines (including weekly training distance and number of training days), physiological data (like VO2max), and anthropometric data (such as BMI and body fat percentage). This multi-faceted approach can help us build a more accurate and comprehensive prediction model.
2. **Prediction Model Development**: While the original study used regression analysis to develop its predictive formula, it opened the door for us to explore both traditional statistical methods and machine learning techniques like k-Nearest Neighbors (kNN) or Naive Bayes for our project. We can compare these methods to determine which offers the most accurate predictions based on our collected data.
3. **Practical Tool Creation**: The ultimate goal of our project, similar to the study, is to create a practical tool that can be used by recreational runners and professionals to set training goals and improve performance. By identifying key predictors of race performance, we can offer personalized advice to runners on how to optimize their training and nutrition.
4. **Technique Application**: Although the study did not use kNN or Naive Bayes, our project could explore these machine learning techniques as innovative approaches to analyze the data. This could be especially useful if we aim to handle larger datasets or wish to experiment with predictive analytics in sports performance more broadly.
5. **Scientific Support for Recreational Athletes**: Just as the study highlighted the need for scientific support for "novice" runners, our project underscores the importance of data-driven insights to support recreational athletes in achieving their goals, whether it's finishing a race in a target time or improving overall fitness.

In conclusion, the study provides a solid foundation for our project by illustrating the potential of combining various data types to predict half-marathon performance. It encourages us to explore beyond traditional statistical methods and consider machine learning techniques that might offer new insights and more precise predictions. By following a similar methodological approach but also incorporating innovative analysis techniques, our project could contribute valuable tools and knowledge to the running community.

### The third article: Application of K-Nearest Neighbors Variants for Disease Prediction

Introduction

Disease prediction poses an increasing challenge in the medical field, with machine learning algorithms being widely utilized to address this issue. Among these algorithms, the k-Nearest Neighbors (kNN) algorithm stands out for its simplicity and adaptability. A recent study published in *Scientific Reports* by Shahadat Uddin et al. (2022), titled "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," explores various kNN variants and their comparative performance in the context of disease prediction.

Key Points of the Research

* The study focuses on evaluating the classic kNN and its variants, including Adaptive KNN (A-KNN), Fuzzy KNN (F-KNN), and Hassanat KNN (H-KNN), among others, to assess their effectiveness in disease prediction.
* The variants were tested using eight benchmark datasets from public sources like Kaggle, UCI Machine Learning Repository, and OpenML, representing various disease contexts.
* Performance metrics such as accuracy, precision, and recall were utilized for comparative analysis, with Hassanat KNN showing the highest average accuracy (83.62%), followed by Ensemble Approach KNN (82.34%).

Discussion and Conclusions

* The study identified Hassanat KNN as the best-performing variant based on accuracy, strongly recommending its application in training programs and competitions.
* This research offers valuable insights for healthcare researchers and stakeholders in selecting the most suitable kNN variant for predictive disease risk analytics.

**Literature Review Paragraphs**

In the study "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction" by Shahadat Uddin et al., an extensive evaluation of various kNN variants and their application in disease prediction is presented. The study leverages multiple benchmark datasets related to different diseases and analyzes the performance of these kNN variants based on accuracy, precision, and recall. The notable finding that Hassanat KNN offers the highest average accuracy points to the potential of this variant for improved disease prediction models. Furthermore, through the thorough comparison between the variants, this research lays the groundwork for healthcare researchers to make an informed choice in selecting the most suitable kNN variant. This contributes to the growing domain of machine learning in medical science, where precision in disease prediction is crucial for both preventive and reactive healthcare strategies.

The insights gained from this study highlight not only the value of machine learning in contemporary medical practice but also challenge further innovation in algorithm development for even more accurate disease predictions. Thus, the research by Uddin et al. serves as a valuable resource for future exploration within this dynamic and impactful field.

This in-depth analysis of kNN variants for disease prediction underscores the crucial role of advanced algorithms in enhancing diagnostic accuracy and treatment efficiency, contributing to more robust and informed medical decision-making.

Afbeelding met tekst, diagram, schermopname, tekenfilm

Automatisch gegenereerde beschrijvingThis visual is an illustration of the k-Nearest Neighbors (kNN) algorithm in action. The kNN algorithm is a simple, yet powerful machine learning technique used for classification tasks (Uddin et al., 2022). Here’s how the diagram relates to the kNN algorithm:

1. **Data Points:** The squares (□) and triangles (△) represent two different classes within a dataset. For instance, in a disease prediction model, one could represent a 'disease' class and the other 'no disease'.
2. **Queries:** The circles (○ ●) represent query points—these are the new data points that we want to classify based on the available data.
3. **Classification Process:** For Query-A and Query-B, the algorithm identifies the 'k' nearest data points (the circles with numbers inside them denote the count of nearest neighbors considered). Based on the majority class of these nearest neighbors, the query point is classified. For example, if 'k' is set to 3, Query-A would be classified in the same category as its three nearest neighbors (two squares and one triangle), and if most of the three closest are squares, then Query-A would be classified as a square (Uddin et al., 2022).
4. **Majority Voting:** After identifying the nearest neighbors, a majority vote is taken to decide the class of the query points. In the figure, Query-B is surrounded by five nearest neighbors, and if most of them are triangles, Query-B will be classified as a triangle (Uddin et al., 2022).

The study at hand delves into the intricacies of the k-nearest neighbor (KNN) algorithm, primarily known for its classification prowess in supervised machine learning and particularly instrumental in disease prediction. The classic KNN methodology, while simple, predicates its prediction on the proximity of a query's features to the labeled examples in the training dataset. This proximity is calculated using a set number of nearest neighbors, or 'k', and the classification is determined by majority vote. Despite its simplicity and adaptability, the algorithm has limitations, such as equal distance weighting and noise sensitivity, which various KNN variants seek to address.

**Adaptive KNN (A-KNN):** This variant advances the KNN algorithm by dynamically determining the optimal 'k' for each testing instance. By training the dataset to find a suitable 'k' value within a specific range, the A-KNN variant customizes the neighbor consideration process, potentially enhancing classification accuracy.

**Locally Adaptive KNN (LA-KNN):** The LA-KNN variant employs a discrimination class approach, which adjusts 'k' by analyzing the distribution of majority and minority classes around a query. This approach allows for a nuanced assessment of neighbor influence, improving the representativeness of the classification decision.

**Fuzzy KNN (F-KNN):** F-KNN introduces the concept of membership values for neighbor classes, thus incorporating a degree of uncertainty and partial belonging rather than strict classifications. This can be particularly useful in medical datasets where diagnoses aren't always clear-cut.

**K-Means Clustering-Based KNN (KM-KNN):** By integrating k-means clustering, this variant reduces the dataset to cluster centroids before applying a nearest neighbor classifier. This method effectively simplifies the dataset, potentially improving the speed and focus of the classification process.

**Weight Adjusted KNN (W-KNN):** The W-KNN variant assigns varying weights to neighbors based on their distance, giving closer points greater influence. This addresses the classic limitation where all neighbors are considered equally, regardless of their distance to the query.

**Hassanat Distance KNN (H-KNN):** Opting for a novel distance metric, the H-KNN variant aims to improve classification efficacy by calculating distances using maximum and minimum vector points, which may be more suitable for certain datasets than traditional Euclidean or Manhattan distances.

**Generalized Mean Distance KNN (GMD-KNN):** This variant stands out by using localized mean vectors and generalized mean distance calculations, which could potentially enhance classification by factoring in localized data distributions.

**Mutual KNN (M-KNN):** M-KNN focuses on mutual neighbors to clean the dataset from noisy data points, leading to a more streamlined and presumably more accurate classification.

**Ensemble Approach KNN (EA-KNN):** EA-KNN eliminates the need to specify 'k' by employing an ensemble approach that iteratively applies the classification process over a range of 'k' values, aggregating the results for a final decision.

The exploration of these KNN variants is rooted in the quest to optimize classification accuracy in disease prediction. Each variant, with its distinctive design and rationale, reflects a unique angle on the classic algorithm's framework, attempting to enhance the predictive quality in complex, real-world scenarios such as medical diagnostics.

Incorporating these KNN variants into a comparative performance analysis offers the potential to uncover the most effective methods for disease prediction. This study, through meticulous experimentation with various disease datasets, seeks to evaluate these variants side by side, thereby illuminating their individual and collective merits. By considering the parameter values that yield optimal performance, the research aims to provide a robust foundation for selecting the most suitable KNN variant for a given predictive task in healthcare. Such an endeavor is not only academically enriching but also pivotal for advancing medical machine learning applications toward more precise and reliable diagnostic tools.

The **Adaptive KNN (A-KNN)** method is recommended for this research. This method dynamically determines the optimal 'k' value for each test point, allowing it to adapt to variations in the dataset and the specific characteristics of the runners and the race. By doing so, it can achieve more accurate predictions that consider individual differences and changes in the race environment. Furthermore, it provides the opportunity to tailor performance predictions to the specific context of each runner and race, which is essential for achieving personalized insights and advice for performance improvement. Through the utilization of data from marathon races over the past five years, the goal of developing a model that predicts a runner's finishing time based on their age, gender, and previous race performance can be achieved. Ultimately, this approach aims to provide runners and coaches with valuable insights and recommendations to enhance their performance.

# 3. Data understanding

For the relevance of the model, it is important to understand the data and what does the data means. This chapter focuses on identifying and analyzing the variables relevant in calculating marathon times. The variables were determined from the literature review. By examining the multifaceted nature of marathon running, we can isolate the key factors that significantly influence an athlete's performance.

**Identifying relevant variables**

Marathon performance is influenced by a constellation of factors ranging from physiological characteristics to training habits. To construct a predictive model with practical relevance and scientific accuracy, it is important to consider variables that capture the essence of an athlete's ability and preparation. Here are the key variables that were deemed relevant to our study:

1. **Physiological Factors**:
   * **Age and Gender**: Age and gender have been consistently shown to impact marathon times, with performance typically peaking at certain ages and differences observed between male and female athletes.
   * **Body Mass Index (BMI)**: While not a direct indicator of fitness, BMI can influence endurance and running efficiency, making it a variable worth considering.
   * **VO2 Max**: Representing the maximum oxygen uptake during intense exercise, VO2 max is a cornerstone measure of cardiovascular fitness and endurance capability.
   * **Lactate Threshold**: The intensity at which lactate begins to accumulate in the bloodstream, indicating the shift to more anaerobic metabolism, is crucial for endurance sports like marathoning.
2. **Training Variables**:
   * **Weekly Mileage**: The amount of distance covered in training per week is a strong indicator of endurance capacity.
   * **Long Run Length**: The maximum distance covered in a single training run can reflect the athlete's endurance preparation for the marathon distance.
   * **Speed Work and Interval Training**: Sessions focused on high-intensity intervals contribute to both aerobic and anaerobic conditioning, essential for marathon pacing strategies.
3. **Previous Performance Metrics**:
   * **Past Marathon Times**: Previous marathon performances provide a baseline for predicting future marathon times.
   * **Shorter Distance Times (e.g., 10K, Half-Marathon)**: Times in shorter races can help estimate marathon potential through established performance prediction models.
4. **Environmental and Course Factors**:
   * **Weather Conditions**: Temperature, humidity, and wind can significantly affect marathon performance on race day.
   * **Course Elevation and Profile**: The topography of the marathon course (e.g., flat, hilly) plays a critical role in the race strategy and performance outcomes.

Analyzing the Impact of Variables

With the relevant variables identified, the subsequent step involves analyzing their impact on marathon performance. Statistical analysis and data exploration techniques will be employed to discern patterns, correlations, and potential causations within our dataset. This analysis aims to quantify the significance of each variable about marathon times, providing a data-driven foundation for feature selection in our machine learning models.

By meticulously selecting and examining these variables, we position ourselves to craft a predictive model that not only forecasts marathon performance with precision but also offers insights into the optimization of training and race strategies. The subsequent chapters will delve into the preparation of our dataset and the development of our machine-learning models, building upon the comprehensive understanding established in this chapter.

# 4. Data Preparation

We initially focused on a comprehensive dataset spanning the marathon results from the years 2015, 2016, and 2017. This dataset, which we refer to as Dataset 1, served as our primary source for developing and testing our predictive models. The subsequent sections detail our approach to preparing this dataset and provide insights into the preliminary model performance, highlighting the challenges we encountered regarding prediction accuracy.

Initial Steps with Dataset 1

Our first task involved a rigorous process of cleaning, consolidating, and preprocessing the data from the three consecutive years. By combining these datasets, we aimed to leverage a broader spectrum of data to enhance the robustness and reliability of our predictive models. The steps taken ensured that our dataset was devoid of inconsistencies, missing values were appropriately handled, and outliers were identified and corrected. Furthermore, feature selection was informed by careful consideration of variables directly influencing marathon performance, as previously outlined.

Normalization and Model Development

Normalization of the features within Dataset 1 it is important to standardize the data and ensure that each variable contributed proportionately to the predictive models. This process was especially crucial given the varied scales of the data points, from demographic information to performance metrics. Following normalization, we embarked on developing our initial predictive models, employing both Artificial Neural Networks (ANN) and K-Nearest Neighbors (KNN) to forecast marathon finishing times.

Encountering Prediction Challenges

Upon evaluating our initial models using Dataset 1, we encountered a significant challenge. The mean squared error (MSE) between the predicted marathon times and the actual finishing times was approximately 12 minutes. This discrepancy highlighted a considerable gap between our model's predictions and the real-world performances of marathon runners. While the models were successful in capturing the general trends and relationships within the data, this 12-minute error margin indicated a need for further refinement and investigation into additional factors that could improve the model's accuracy.

Reflections and Next Steps

The initial findings from our work with Dataset 1 underscore the complexity of accurately predicting marathon performance. Several factors may contribute to the observed 12-minute mean squared error, including the need for more sophisticated feature engineering, the incorporation of additional relevant variables do not present in Dataset 1, or the exploration of more complex model architectures.

The insights gained from this initial phase of our project are invaluable. They provide a clear indication that while we have made significant progress, there is room for improvement in our models. As we move forward, our focus will shift towards addressing these challenges. This may involve revisiting our data preparation steps to include more granular or context-specific variables, experimenting with different machine learning algorithms, or employing advanced techniques such as ensemble methods or deep learning to enhance our models' predictive power.

In summary, our journey with Dataset 1 has laid a foundational understanding of the task at hand and illuminated the path forward. By recognizing the limitations of our initial models and identifying the mean squared error as a key area for improvement, we are better positioned to refine our approach and strive for a more accurate prediction of marathon performance in subsequent phases of our project.

# Model Development and Training

This chapter delves into the advancements made in the modeling phase, focusing on the evaluation of the model's performance through key statistical measures: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared Score.

## 5.1. Model Refinement Strategies

In response to the initial challenges encountered with Dataset 1, we employed a series of model refinement strategies aimed at reducing prediction errors and improving the model's explanatory power. These strategies included:

* **Enhanced Feature Engineering**: Diving deeper into the data, we extracted more nuanced features that could potentially influence marathon performance, such as pacing strategies and training consistency.
* **Algorithm Optimization**: We experimented with various hyperparameters for both the ANN and KNN models, seeking the optimal configurations that minimized prediction errors.
* **Incorporation of Advanced Techniques**: Recognizing the complexity of marathon performance prediction, we explored advanced modeling techniques such as ensemble methods and deep learning to capture more intricate patterns in the data.

Evaluation of Model Performance

The culmination of our refinement efforts was evaluated through the following key performance metrics:

* **Mean Squared Error (MSE): 521,207.3630973528**
  + The MSE offered insight into the average squared difference between the actual and predicted finishing times. Although a lower MSE is desirable, its interpretation is somewhat abstract due to the squared units. The value indicated that while our model was capturing the general trends, there was still a notable variance in prediction accuracy, especially for larger errors.
* **Mean Absolute Error (MAE): 473.05386345166096**
  + The MAE provided a more intuitive measure of the model's average prediction error, amounting to approximately 473 seconds or roughly 7.88 minutes off from the actual marathon times. This metric was particularly useful for understanding the model's performance in units directly relatable to marathon times.
* **R-squared Score: 0.915532329936932**
  + The R-squared score, representing the proportion of variance in marathon times explained by the model, stood at an impressive 91.5%. This high value indicated that our refined model was highly effective in capturing the relationship between the selected features and marathon finishing times.

Interpretation and Implications

Our refined model demonstrated a significant improvement in predicting marathon finishing times, with an average prediction error of about 8 minutes. Despite the inherent challenges in forecasting such a complex outcome, the high R-squared score suggested that our model succeeded in identifying and leveraging the critical patterns within the data.

The insights gleaned from the MSE and MAE further underscored the importance of continual model optimization. While the current model performed admirably, the presence of prediction errors highlighted opportunities for further enhancements, possibly through the integration of additional contextual factors or the exploration of more sophisticated modeling techniques.

## 5.2.Training the model and playing with the variables

In the development of our prediction model, the K-Nearest Neighbours (KNN) algorithm played a key role in the prediction of marathon finish times. The training process unfolded in several structured steps to ensure the robustness and reliability of the model.

First, we identified and selected relevant features, including '10K\_seconds', 'Age', 'Gender' and 'Half\_seconds' as predictor variables, with 'Official\_Time\_seconds' as the target outcome. To ensure consistency and comparability across these features, we applied Standards Caler, a pre-processing step that standardized the data by normalizing feature values.

The dataset was then divided into training and test subsets using an 80-20 split. This division allowed us to train the model on a substantial portion of the data, while reserving a separate set for evaluating its predictive accuracy.

After configuring the training set, we instantiated the KNN model, specifying five Neighbours (n\_neighbors=5) to be considered in the prediction process. This parameter was chosen to balance the trade-off between overfitting and underfitting, aiming for a model that generalizes well to unseen data.

The training phase for the KNN model was straightforward, involving the storage of the feature vectors of the training data and the corresponding target values. Unlike models that undergo extensive parameter tuning during training, KNN's 'training' consists primarily of memorizing the training data set. This characteristic underlines the classification of KNN as a type of "lazy" learning algorithm, where the computational intensity is postponed until the prediction phase.

During prediction, the model calculates the distances between a new data point and all the points in the training set to identify the nearest neighbours. The marathon finish time is then predicted by averaging the target values of these nearest neighbours, which is the essence of regression tasks.

Our model for predicting marathon finishing times initially had a mean absolute error (MAE) of 12 minutes, using primarily the 10km split time as the key variable for prediction. To improve the accuracy of the model, we focused on reducing the MAE by including half marathon times, a strategic decision based on the assumption that longer distance performances might be a more reliable indicator of marathon potential.

Further improvement was achieved by adding key variables identified as significant in the literature: age and gender. These additions have been instrumental in refining our understanding of the factors that influence marathon performance, leading to a reduction in the MAE to 8 minutes. At this point, our progress temporarily stalled as we considered the next steps to further strengthen the reliability of the model.

A breakthrough was achieved by addressing outliers in both the 10km and half marathon times. Removing these outliers was a critical step in our analysis, sharpening the model's focus on the most representative data. This adjustment further reduced the MAE to 7.5 minutes, a significant step towards improving the accuracy of our marathon time predictions.

This process not only demonstrates the nature of model refinement, but also highlights the importance of integrating diverse data points and methodically addressing data quality issues to improve predictive accuracy. As we continue to refine our model, these steps underscore the ongoing journey to improve its reliability and validity in predicting marathon finish times.

## 5.3. Development and Evaluation of the Neural Network Regression Model

In developing our predictive model, we used the TensorFlow Keras API to construct a neural network designed for regression tasks. The construction and refinement of the model was carried out through several key steps, which are outlined below:

**Model Definition**

Our neural network model is structured as a sequential model, a choice motivated by its straightforward linear stack of layers, which is well suited to the requirements of our problem. The architecture consists of three dense layers:

* Input Layer:

The initial layer has 64 neurons and uses the Rectified Linear Unit (ReLU) activation function. The input\_shape parameter is specified according to the dimensionality of our input features, ensuring that the model accurately interprets the structure of the input data.

* Hidden layer:

Following the input layer, another dense layer with 64 neurons and ReLU activation is added, enhancing the model's ability to capture complex relationships in the data.

* Output Layer:

The network concludes with a single neuron dense layer. This layer does not use an activation function as it is designed to output a continuous value, in line with the regression nature of our task.

**Model construction**

Once the model is defined, it is compiled using the Adam optimiser - an algorithm known for its efficient computation and adaptability to different data types and architectures. The mean squared error (MSE) is chosen as the loss function, aiming to minimise the mean squared difference between predicted and actual values, which is a standard approach in regression analysis.

**Model training**

The training process is performed over 100 epochs with a batch size of 32. This setup implies that the model weights are adjusted after processing every 32 samples, iterating through the entire dataset 100 times.

# 6. Evaluation and Iteration

Comparing our predictive model to the model described in the provided article on marathon performance prediction using artificial intelligence (AI) techniques, we observe several similarities and key differences in approaches, methodologies, and outcomes. Both endeavors aim to harness the power of machine learning to forecast marathon finishing times with a high degree of accuracy, leveraging runner characteristics and race performance data.

## 6.1. Similarities

1. **Data Utilization**: Both models employ marathon results data spanning multiple years, using this information to predict finishing times based on various runner attributes, including age, gender, and partial race times (such as 10K or half-marathon times).
2. **Feature Engineering**: In both cases, significant emphasis is placed on transforming the data into a machine-readable format, particularly converting race times from strings to numerical values (seconds) to facilitate model computation.
3. **Inclusion of Key Variables**: Age, gender, and earlier race segment times (e.g., 10K and half-marathon times) are recognized as crucial predictors of marathon performance, featuring prominently in both modeling efforts.

**Differences**

1. **Model Complexity and Variety**:
   * Our approach is characterized by the incremental development of predictive models, starting from simple linear regression to more complex models like K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN). This progression allows for a nuanced exploration of how different models capture the relationship between predictors and marathon times.
   * The article's model, while similar in using AI techniques, does not detail a progression through varying levels of model complexity or the exploration of different machine learning algorithms beyond ANN and KNN.
2. **Evaluation Metrics**:
   * Our model's performance is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared scores, offering a comprehensive view of prediction accuracy and the model's explanatory power.
   * The comparison with the article's model on these specific metrics is not direct, as the article focuses more on the conceptual application of ANN and KNN rather than detailed statistical evaluation metrics.
3. **Model Refinement and Outcomes**:
   * Our model underwent significant refinement, evidenced by the improvement in predictive accuracy as reflected in the decrease of MSE and MAE, and an impressive R-squared score indicating that approximately 91.5% of the variance in marathon times is explained by our model.
   * The model from the article, while validated using ANN and KNN to predict marathon performance accurately, does not specify the extent of error reduction or improvement in fit achieved through model refinement processes similar to ours.
4. **Implementation Details**:
   * Our workflow is described as a step-by-step Python script that includes data preparation, linear regression, KNN, and ANN implementation, providing a detailed roadmap for replication and understanding.
   * The article, while discussing the use of ANN and KNN, focuses more on the theoretical application of these methods without delving into the specifics of the coding implementation or the iterative process of model improvement.

**Conclusion**

In comparing our model to that described in the article, it's evident that while both approaches aim to leverage AI for marathon performance prediction, our methodology offers a more granular look at model development, evaluation, and refinement. Our use of a diverse set of evaluation metrics and a step-by-step Python script enhances the transparency and reproducibility of the predictive modeling process. This not only demonstrates the model's efficacy in capturing the nuances of marathon performance prediction but also provides a clear framework for further exploration and improvement in the field of sports analytics and machine learning.

## 6.2. Assessment of Relevant Variables from Theoretical Framework

Our theoretical framework identified a set of variables critical to predicting marathon performance, categorized into physiological factors, training variables, previous performance metrics, and environmental and course factors. Here in, we assess how effectively these variables have been integrated into our model:

* **Physiological Factors (Age, Gender, BMI, VO2 Max, Lactate Threshold)**: We acknowledged the significant impact of age and gender on marathon times, incorporating these as fundamental features in our model. It is important to add variables such as BMI and the more complex physiological measures of VO2 Max and Lactate Threshold were challenging to incorporate due to data availability constraints. Their potential predictive power underscores an area for future data enrichment. It would be recommended to add it to the model to predict a better marathon time.
* **Training Variables (Weekly Mileage, Long Run Length, Speed Work, and Interval Training)**: It would be wise to add to our model weekly mileage and long run length as direct indicators of an athlete’s endurance training status. However, quantifying the impact of speed work and interval training required innovative feature engineering to translate these qualitative aspects into measurable inputs for our model.
* **Previous Performance Metrics (Past Marathon Times, Shorter Distance Times)**: Leveraging historical performance data, including past marathon times and shorter distance race times, proved to be a robust predictor within our model. This approach allowed us to anchor predictions in empirical evidence of an athlete’s running capabilities.
* **Environmental and Course Factors (Weather Conditions, Course Elevation and Profile)**: These variables posed a unique challenge due to their variability and external nature. While we made efforts to incorporate weather conditions and course profiles into our model, the complexity of accurately modeling their impact on performance highlighted the need for sophisticated data collection and modeling strategies.

Utilization of Data in Model Development

Our approach to data preparation and feature engineering aimed to capture the multifaceted nature of marathon performance. By meticulously cleaning our dataset and deriving new features from the identified variables, we ensured a dataset to train our model. The transformation of raw race times into computable formats and the categorization of qualitative training data into quantifiable metrics were key steps in aligning our model with theoretical insights.

Model Validity Description

The validity of our model is primarily assessed through its predictive accuracy and the ability to generalize across different datasets. The inclusion of diverse and theoretically grounded variables has endowed our model with a strong foundation for capturing the nuances of marathon performance. However, the mean squared error (MSE) and mean absolute error (MAE) metrics revealed areas for improvement in model precision.

The model's R-squared value, indicating that a significant portion of the variance in marathon finishing times can be explained by our model's inputs, attests to its effectiveness. Yet, the recognition of areas where our data and variable utilization can be enhanced suggests pathways for increasing model validity.

* **Future Enhancements for Model Validity**: To augment the validity of our model, we aim to further refine our feature selection and engineering processes, incorporate more complex physiological and environmental variables as data becomes available, and continuously validate our model against new and diverse datasets.

In conclusion, our evaluation of the relevant variables against our data and model is the basis for improvement in our predictive approach. As we continue to refine our model, our focus remains on enhancing its validity, ensuring it remains a reliable tool for athletes, coaches, and sports scientists in the planning and optimization of marathon training and performance strategies. It is important to find out more about the variables to make our model more predictable and more validated. We trained our model from a score of 12 minutes on the MSE to 8 minutes.

## 6.4. Naive Bayes

Naive Bayes (NB) theory is based on Bayesian statistics and provides a powerful yet simple probabilistic approach to classification tasks. It is particularly well known for its effectiveness in text classification, including spam detection and sentiment analysis, although its applications span many domains. The essence of Naive Bayes lies in its use of probability to make predictions, assuming that the presence (or absence) of a particular feature in a class is unrelated to the presence (or absence) of any other feature.

### 6.4.1. Naive Bayes Theory

Core principles

Bayes' Theorem: At the heart of NB theory is Bayes' theorem, a fundamental principle of probability theory that describes the probability of an event given prior knowledge of conditions that may be associated with the event. The theorem is expressed mathematically as:

*P* (*A*∣*B*) = *P*(*B*)*P*(*B*∣*A*) ⋅ *P*(*A*)

where:

* *P*(*A*∣*B*) is the probability of event A given that B is true.
* *P*(*B*∣*A*) is the probability of event B given that A is true.
* *P*(*A*) and *P*(*B*) are the probabilities of observing A and B independently of each other.

**Naive assumption:** The 'naive' aspect of Naive Bayes comes from the assumption that all predictors (or features) are independent of each other, given the class. This simplification makes the model much more tractable, especially for datasets with a large number of features. Despite its simplicity, this assumption often holds surprisingly well for many real-world problems.

Applications and Advantages

Naive Bayes classifiers are especially favored for their efficiency and simplicity, making them suitable for:

* High-dimensional datasets.
* Real-time predictions.
* Text classification tasks, where they can handle large feature spaces with ease.

Among the advantages of Naive Bayes are its ability to handle an immense number of features and its straightforward implementation, which together contribute to its widespread use in both academic and industrial settings.

Limitations

While Naive Bayes is powerful, it is not without limitations. The assumption of independent features can sometimes lead to inaccuracies, especially in cases where the relationship between features is significant. Additionally, its performance can be affected by data scarcity for any given feature-class combination, leading to probability estimates that may not reflect the true distributions.

### 6.4.2. Limitations of Naive Bayes in Marathon Time Prediction

While Naive Bayes (NB) theory provides a robust framework for classification tasks, its application to predicting marathon times - a regression problem - presents several challenges, especially when dealing with complex datasets. The main reasons for its unsuitability in this context lie in the basic assumptions and limitations of the NB approach:

**Predictive nature**

**Classification vs. regression:** NB is inherently a classification algorithm, designed to assign category labels to instances based on the likelihood of features belonging to each category. However, marathon time prediction is a regression problem that aims to predict a continuous outcome (the finish time) based on various predictors. This fundamental mismatch makes NB an inappropriate tool for directly predicting marathon times.

**Feature independence**

**Interdependent features:** One of the core assumptions of NB is the independence of features given the class label. However, in predicting marathon performance, many features are likely to be interrelated. For example, training intensity, previous performance metrics, and physiological factors such as VO2 max and lactate threshold are all interrelated and collectively contribute to an athlete's marathon time. The naive assumption of independence between these variables can oversimplify the model and lead to inaccurate predictions.

**Data complexity**

**Complex relationships and interactions:** Marathon performance is influenced by a myriad of factors including training history, nutrition, injury history, race day weather conditions and even psychological state. The simplistic approach of the NB model may not adequately capture the complex, non-linear relationships and interactions between these variables.

**Data continuity**

**Handling continuous data:** NB typically requires data to be categorical or discretised when dealing with continuous variables. While there are variants of NB that can handle continuous data (such as Gaussian NB), the process of discretising or making assumptions about the distribution of continuous variables (such as marathon times or physiological measurements) can lead to loss of information and precision.

In summary, although Naive Bayes is powerful for certain types of classification problems, it is not ideally suited to the task of predicting marathon finish times. We have used the theory, and it is in our Github document, but the MAE shows a big difference with using the KNN theory. The complexity of the dataset, the interdependence of the predictive features and the continuous nature of the outcome suggest that a more sophisticated regression model or machine learning approach capable of capturing complex relationships and interactions would be more appropriate and likely to yield more accurate and reliable predictions.

# 7. Deployment

This chapter outlines how we intend to disseminate the results and the practical applications of our developed models, detailing our findings and offering recommendations for future research and implementation in training programs and competitions.

Presentation of Results

Our findings, from data collection through to model refinement and validation, are summarized in this report.

* **Report**: The report will serve as a detailed record of our research, offering a thorough examination of our data analysis, model development, and evaluation process. It will include an in-depth discussion of the theoretical underpinnings of our variable selection, the statistical significance of our findings, and the implications for marathon runners and coaches.

Recommendations for Future Research

Building upon the groundwork laid by our project, we propose several avenues for future research that could further enhance the predictive modeling of marathon performance:

* **Data Enrichment**: Future studies should aim to incorporate more data on training routines, physiological measures, and environmental conditions to enrich the model’s input variables.
* **Advanced Modeling Techniques**: Exploring more machine learning algorithms and deep learning architectures may uncover new insights and improve predictive accuracy.
* **Real-World Application and Validation**: Implementing the model in real-world scenarios, such as personalized training programs or race strategy optimization, could provide valuable feedback for model refinement.

Application in Training Programs and Competitions

The practical application of our models extends beyond academic interest, offering tangible benefits for the running community:

* **Personalized Training Plans**: By inputting an athlete's specific characteristics and training data, coaches can use our model to tailor training plans that optimize performance outcomes.
* **Race Strategy Development**: Athletes and coaches can apply predictions from our model to strategize race pacing and energy distribution, potentially improving race day performance.
* **Competitive Analysis**: Teams and sports organizations can leverage the model to analyze competitors' performances, aiding in strategic planning for upcoming competitions.