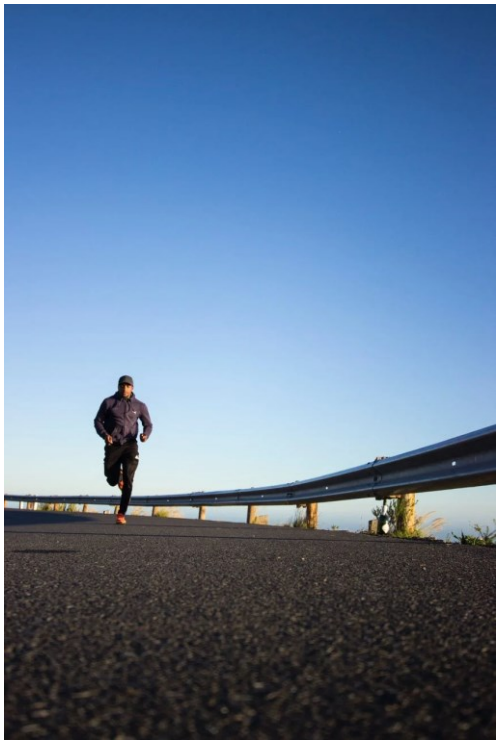

PREDICTING WATER STATIONS AT THE ENSCHEDE MARATHON

Master's Thesis Industrial Engineering & Management

ABSTRACT

A study about the feasibility of using machine learning to better predict the location of water stations at running events like the Enschede Marathon.

Author

Albert Simons

Supervisors University of Twente

Dr. I. Seyran Topan
Dr. A. Karahanoglu
Dr. E Topan

Supervisor Enschede Marathon

S. Melief

Acknowledgements

I would like to express my deepest gratitude to the following individuals who have played a crucial role in the completion of my thesis and my master's journey. Their unwavering support, guidance, and encouragement have been invaluable to me.

First and foremost, I would like to thank my girlfriend Nienke. Throughout the entire process, she stood by my side, offering her unwavering support and understanding. She gave me the strength to overcome challenges and persevere. I am truly grateful for her constant presence and the patience she showed me during a period where I might not always have been the greatest person to be with.

I am indebted to my father for his ability to strike the delicate balance between challenging me with tough questions and not pushing me away. His insightful inquiries pushed me to keep on going and, I think, are the most important reason that I could finish this thesis, and the rest of my study, as it is now.

To my mother, I owe a debt of gratitude for her unwavering support and encouragement. Her belief in my abilities and her reassurance during moments of self-doubt were instrumental in keeping me motivated and focused throughout my master's journey.

My brother deserves a special mention for his vast knowledge about seemingly, and sometimes annoyingly, everything. His diverse insights and expertise in various areas provided me with valuable perspectives and helped me refine my research approach, especially during the early stages.

I am immensely grateful to my supervisors, Ipek, Engin, and Armagan. Ipek invested significant effort in both the supervision of my thesis and personal guidance. Her insightful questioning about my progress and well-being ensured I stayed on track and continuously improved. Engin's profound understanding of Machine Learning and his valuable insights guided me in applying the appropriate techniques in my research. Armagan's expertise in the field of marathon running brought a unique perspective to my thesis, providing invaluable insights that enriched my work.

Lastly, I would like to extend my gratitude to all the friends and family members who offered their support, encouragement, and understanding throughout this journey.

Completing this thesis and my master's degree would not have been possible without the unwavering support and assistance of these incredible individuals. I am truly fortunate to have had their guidance and encouragement.

Management Summary

The Enschede Marathon is a popular annual running event held in the Netherlands. The amount and placement of water stations is an important factor for the successful completion of an endurance event. The current placement of water stations at the Enschede Marathon is based on experience and has difficulty in correctly taking into account the impact of multiple environmental factors such as temperature, wind, humidity, and elevation. With the increasingly extreme weather conditions due to climate change, there is a need to investigate the placement and frequency of water stations.

The main research question in this study is "How can Enschede Marathon use a data-driven method to find the optimal placement frequency for water stations to achieve a sufficient performance of runners?" The research is divided into three sections: understanding the current situation at Enschede Marathon, reviewing the literature on water management at endurance sports events and data-driven methods, developing a model, and evaluating model performance. The sub-questions include understanding the current organization and features considered for water stations at Enschede Marathon, identifying conditions to predict water station locations, data collection practices, data preparation for modeling, selecting and comparing models, identifying the best-performing method, evaluating model performance, identifying significant features, and discussing the implications for Enschede Marathon. To address these questions, this study will review the literature and analyze data from the Amsterdam, Rotterdam, and Lisbon Marathons to predict the influence of several features on water stations. Based on the findings, recommendations will be made to optimize the placement and frequency of water stations at the Enschede Marathon to ensure the runners' safety and performance.

This study compared five machine learning models: decision tree, gradient-boosting trees, random forest, linear regression, and artificial neural network. It concluded that gradient-boosting trees performed best in predicting the pace, as pace best indicates the performance of the runners. The performance of this algorithm was good, but the learning curve indicated possible problems with the data. The results confirmed that temperature and humidity had the most significant impact on the expected pace. An increase in temperature led to an increase in pace, whereas an increase in humidity led to a decrease in pace. The combined effect of temperature and humidity showed that an increase in humidity would decrease the pace, especially at higher temperatures. As shown in Figure I, water stations significantly impacted the expected pace, with a decrease in the pace of up

to nine water stations.

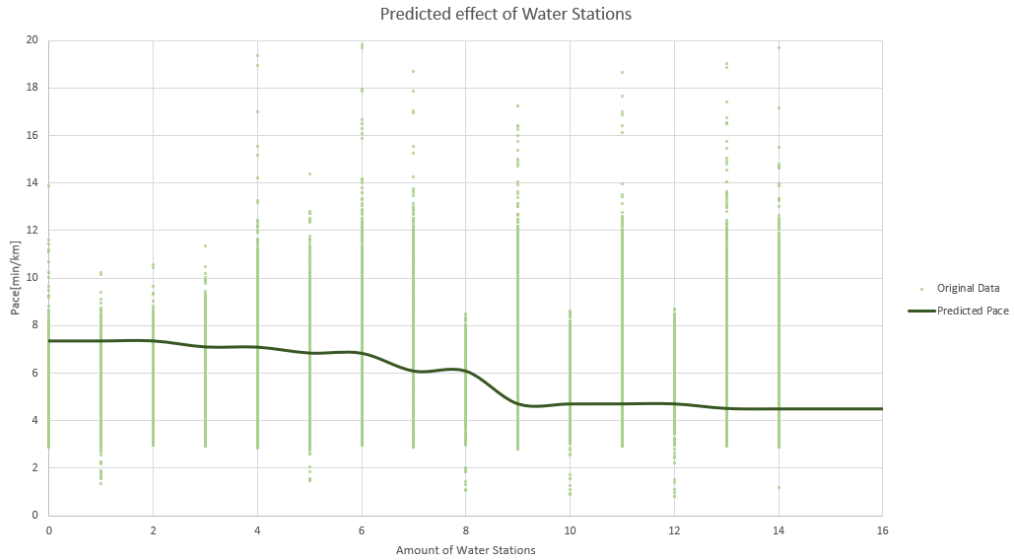

FIGURE I: PREDICTED EFFECT OF WATER STATIONS

It was advised that water stations should be placed evenly over the course of a marathon. Several scenarios were created to show the results for situations with extreme weather conditions, which showed that the organization of the Enschede Marathon should place between 6 and 9 water stations evenly over the course of the marathon for high temperature and humidity and between 3 and 5 for low temperature and humidity. These four scenarios are summarized in Table I and the results are shown in figure II.

	Temperature	Humidity	Windspeed	Recommended Amount of Water Stations
Scenario 1	Low	Low	High	5
Scenario 2	Low	Low	Low	3
Scenario 3	High	High	High	9
Scenario 4	High	High	Low	6

TABLE I: FOUR SCENARIOS

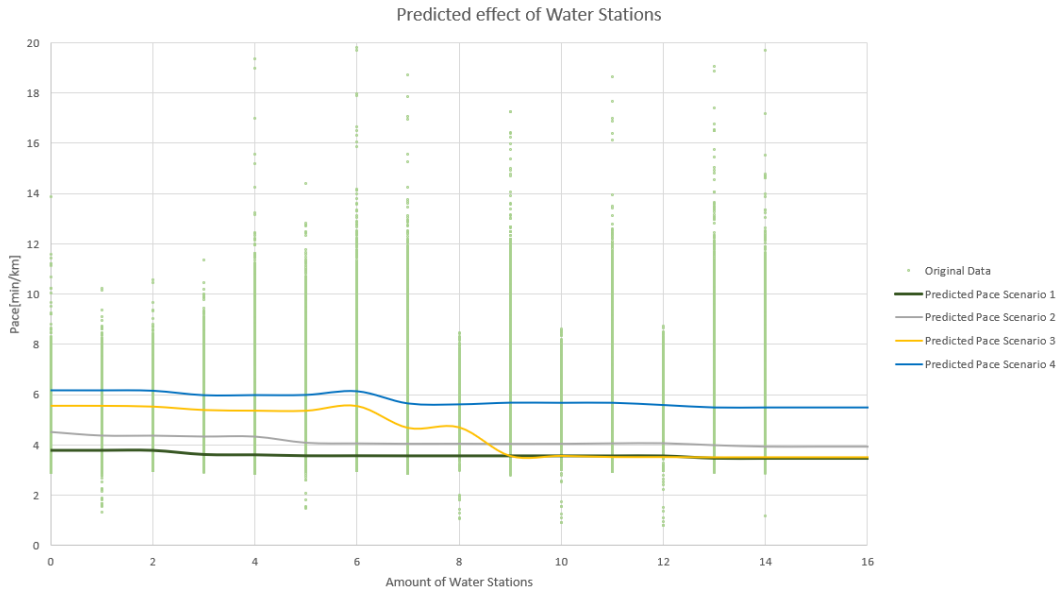

FIGURE II: PREDICTED EFFECT OF WATER STATIONS FOR FOUR SCENARIOS

The recommendation for the Enschede Marathon is to follow these results for future marathons but keep in mind that runners should still be able to drink the recommended 400-800 mL per hour. Future research should include other features, next to the environmental features, in their data like heart rate and sweat loss, and include data from more marathons to improve generalizability.

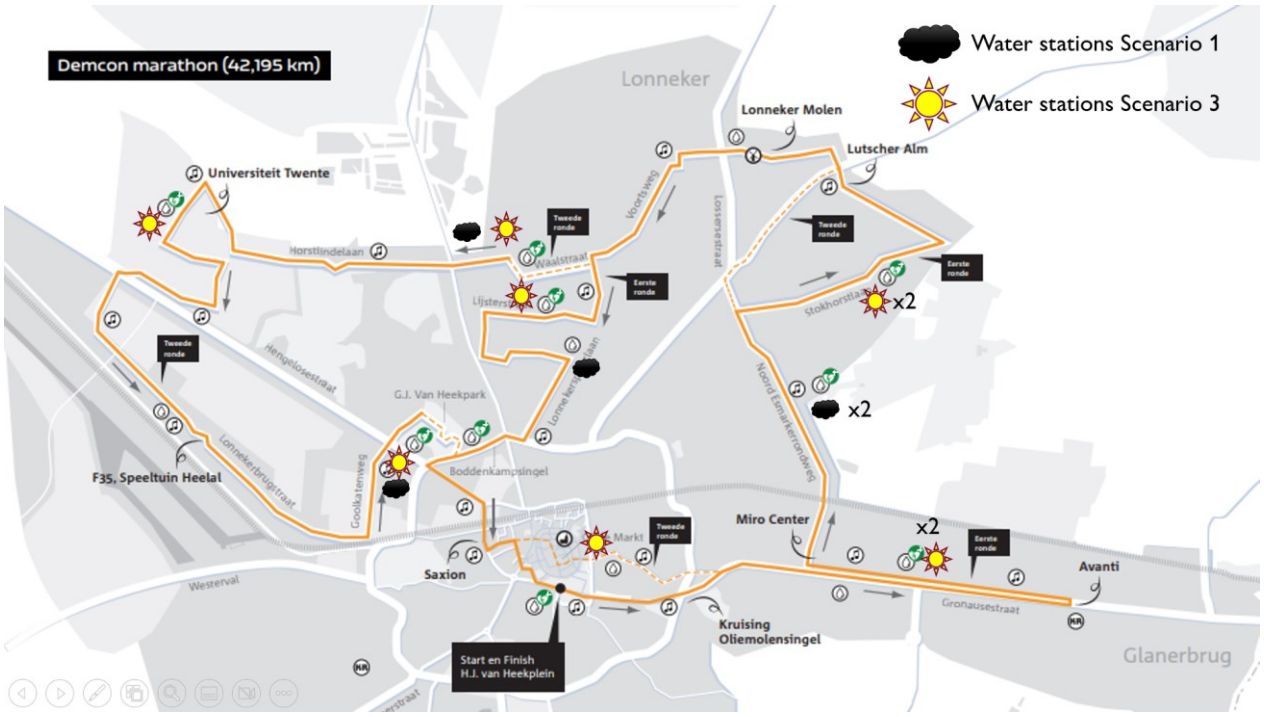

FIGURE III: ENSCHEDE MARATHON WITH TWO SCENARIO'S

Table of Contents

1 Introduction

This study is done in collaboration with the Enschede Marathon. In this chapter, the Enschede Marathon is introduced, the problem statement is described, the research questions are presented and the current situation at the Enschede Marathon is depicted.

1.1 Water consumption at endurance activities

Currently, limited research about the precise placement of water stations, for marathons and other sports events, is available. The amount of research on water intake during these events is substantial. Wyndham and Strydom (1969) state in the late seventies that the tendency at the moment to only drink small quantities of water during a marathon could be dangerous, especially in warm weather. Drinking water is important to compensate for sweat loss which is used for thermal balance in the body (Shapiro et al., 1982). In this light, according to Dancaster and Whereat (1971), runners should consume between 5 and 8 liters of fluids during a marathon. Other researchers make a distinction between elite and regular athletes. Noakes (2003) advises elite athletes to adequately hydrate by ingesting about 200-800 mL/hour. Cheuvront and Haymes (2001) argue that regular athletes should be urged to drink ad libitum (according to the dictates of thirst) with a maximum of 400-800 mL/hour. Both studies point out that excessive drinking could lead to hyponatremia, a low sodium concentration in the blood that could lead to fatalities.

As a runner, it is therefore important to understand the limits and the impact of drinking water during endurance activities. Williams et al. (2012) found that 80% of the runners at the London Marathon perceived that they knew enough about fluid intake. However, only 25% of the runners identified thirst as the most important factor determining their fluid intake. Therefore, not following the "ad libitum" fluid intake strategy is recommended. This led to more than 20% of the runners planning to take water from all 24 water stations. Williams et al. (2012) argue that this could be dangerous because this increases the chance of hyponatremia. Further increase in the amount of water stations could lead to an even higher chance of hyponatremia. These findings suggest a lack of a sufficient understanding of fluid intake strategies among participants in endurance activities, even though most participants claim to know enough. Effective education could provide a solution to this problem, but these findings also stress the importance of the frequency of water stations as this seems to influence the fluid intake strategies.

1.2 Enschede Marathon

The Enschede Marathon is a popular annual running event held in the city of Enschede, Netherlands. This event includes a marathon, half marathon, 10-kilometer, 5-kilometer, and even a 1-kilometer event for kids. The event dates back to July 1947 and is thereby the oldest marathon in Western Europe. It started with 51 runners, but for the 2022 edition, almost 11.000 runners participated in one of the distances (*Historie - Enschede Marathon*, 2022). Over the years the organization also became more professional and it strives for optimal conditions for the runners of its marathon events

Enschede Marathon is interested in using scientific knowledge to improve its events. For this reason, they collaborate with researchers from the University of Twente and take part in "Science meets the runner" events. An example of a topic of one of these events was a presentation during

the 2022 event, where a researcher talked about the effects of running in cold or warm weather. One area that influences the conditions for runners, and is a part of the marathon design, is the amount of water available during the marathon. The reason is that the performance of endurance activities seriously declines when dehydration exceeds 2% of body mass (Casa et al., 2005). The organization of a running event should provide sufficient water stations so that runners have enough opportunities to hydrate enough and avoid dehydration. The organization of the Enschede Marathon believes that weather conditions play a significant role in the amount of water that is needed by the runners. Although the marathon is held in April each year, weather conditions in the Netherland in April can change significantly each year. Weather conditions tend to be even more extreme due to climate change. This made the Enschede Marathon requestioning the frequency of the water stations and at which locations these water stations should be placed.

1.3 The problem statement

As stated, Enschede Marathon is questioning whether its current design for the race is sufficient in the future. Their main concern in this design is the supply of water to the runners in the form of water stations during the race. Increasing extreme weather conditions due to climate change could have a big impact on the water consumption of the runners and thus could influence the placement of water stations. There probably are however other factors that could influence the placement of water stations. It is generally accepted that temperature for example has an impact on the placement of water stations, but what is the effect of the combination of wind, humidity, and elevation?

Currently, the water stations are being placed based on experience. This means that a selection of experienced runners decides how many water stations are needed and at which location they should be placed a few months before the race. Some alterations can still be made just before the race based on temperature, but this method is unsuitable to consider the impact of multiple combined factors. At the root of this lies a lack of use of data from previous years and other running events to predict the influence of several factors. In Figure 1 the problem bundle is visualized. A core problem should be a problem that could be solved or influenced by the problem owner. Climate change does not fill these requirements and thus is not the focus of this research. The problem statement that is the aim of this research is as follows:

There is no data-driven prediction method for the placement of water stations at the Enschede Marathon.

In Figure 1, the core problem is marked in green. A method that is based on data could provide insight into how different factors influence the water station placement, thereby helping find locations for these water stations and possibly helping improve the current race design of the Enschede Marathon. Therefore, this research aims to create a data-driven method and implement this at the Enschede Marathon.

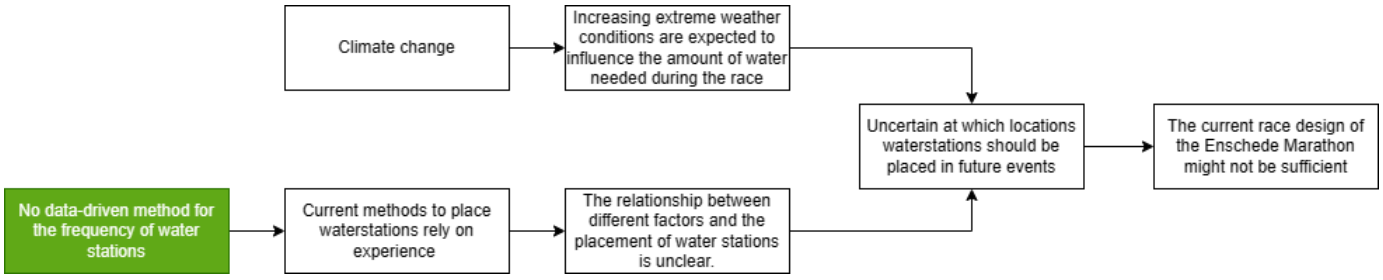

FIGURE 1: PROBLEM BUNDLE

1.4 Research questions

In the process of finding a solution to the core problem, stated in section 1.2, the following research problem has to be solved first: *Enschede Marathon does not know how data could be used to further understand the impact that external factors have on the placement of water stations.* The external factors in this problem could range from environmental factors like temperature to the elevation of the marathon route. To find a solution to the research problem, the following main research question is stated:

How can Enschede Marathon use a data-driven method to find the optimal placement frequency for water stations to achieve a sufficient performance of runners?

Several sub-questions are formulated to answer the research question. These sub-questions are structured in three sections, forming this report's base.

The current situation at Enschede Marathon

Understanding the current water station frequency method at the Enschede Marathon is an important first step in this research. Both the results of this method, the realized water station placement, and the method itself are of importance. As stated, multiple conditions could influence the water station placement. Before continuing to the literature, it is important to understand which of these conditions are currently taken into consideration. The following two research questions are stated to create a better understanding of the current situation.

1. *How are the water stations currently organized at the Enschede Marathon?*
2. *Which features are currently taken into account to predict the location of water stations?*

Literature framework

A framework of findings in the literature is created to answer the main research question. This framework is divided into questions on water management at endurance sports events and questions on the implementation of data-driven methods. This results in the following five questions, which are answered in chapter 2:

3. *Which conditions should be taken into account to predict the location of water stations?*
4. *What data is currently being collected by marathon organizations?*
5. *What data-driven modeling methods are proposed in the literature for solving similar problems?*
6. *How should the data be prepared for the selected methods?*

7. *How can different models be compared and selected for this dataset?*

Model performance

Model performance is a measure of how well a model can make predictions on unseen data. Research on the performance of multiple models is done. Four questions about the best-performing method, the impact of different features, and the implications for the Enschede Marathon are stated below. These questions will be answered in chapter 4 and 5 of this research.

8. *Which method performs best for the given dataset?*
9. *How does the developed model, based on the chosen method, perform?*
10. *Which features have a significant impact on the prediction of the water stations?*
11. *How can the results be used at the Enschede Marathon?*

1.5 Research Outline

The research outline provides an overview of this research project. The main objective of the research is to predict the location of water stations based on various factors. The research consists of six chapters, each of which is described in detail.

The first chapter*, Introduction*, provides background information and motivation for the study. It also outlines the research question and objectives and describes the current placement of water stations at the Enschede Marathon. The second chapter, *Literature Review*, focuses on the features that should be taken into account to predict the frequency of water stations, as well as the data-driven modeling methods proposed in the literature. The third chapter, *Methodology*, outlines the procedures used in the study, including data collection and analysis methods.

The fourth chapter, *Results*, presents the data and analysis that were conducted, including the comparison of the performance of different methods on the dataset, the analysis of the developed model, and the examination of the impact of different conditions on the prediction of water stations. The fifth chapter, *Discussion*, interprets the results in the context of the Enschede Marathon and provides recommendations for future research. The final chapter, *Conclusion*, summarizes the main findings, implications of the study, limitations, and recommendations for future work.

In summary, the research project provides a comprehensive overview of the water station frequency at the Enschede Marathon, including the literature review, methodology, results, discussion, and conclusion. The main goal is to predict the location of water stations based on various factors and to make recommendations for future work.

1.6 Current water stations at the Enschede Marathon

The organization of water stations at the Enschede Marathon is a crucial aspect of the event, as the organization believes proper hydration is essential for the health and performance of runners. For the Enschede Marathon organization, it is important to determine the number and location of water stations based on the length of the course and the expected weather conditions. In general, there should be at least one water station every five kilometers, with additional stations in hot or humid weather. Water stations should also be placed in visible and easily accessible locations, such

as at intersections or near aid stations. Water stations should also be placed in locations that are close to local water taps. Figure 2 shows a blueprint of the map of the 2022 Enschede marathon including the water stations. The placement of these water stations is based on the advice of an experienced panel that uses a combination of experience and the previously stated factors. In 2022, the panel advised more water stations compared to the blueprint, because of the warm and humid weather. Other water stations related factors that are important to the organization are waste management and the provision of other hydration and nutrition options, such as energy drinks and different types of fruit.

The organization and the experienced panel state that there is also a limitation to the number of water stations. One reason is logistical, it can be challenging to set up and staff a large number of water stations along the course. Another reason is safety, having too many water stations can lead to bottlenecks and congestion, which can be dangerous for runners. Additionally, having too many water stations can create unnecessary waste, as excess cups and other materials may not be properly disposed of.

Overall, the effective organization of water stations is a vital part of a successful marathon event. By carefully planning the number, location, and type of hydration options and implementing responsible waste management practices, the marathon organization can ensure that runners have the hydration they need to perform at their best.

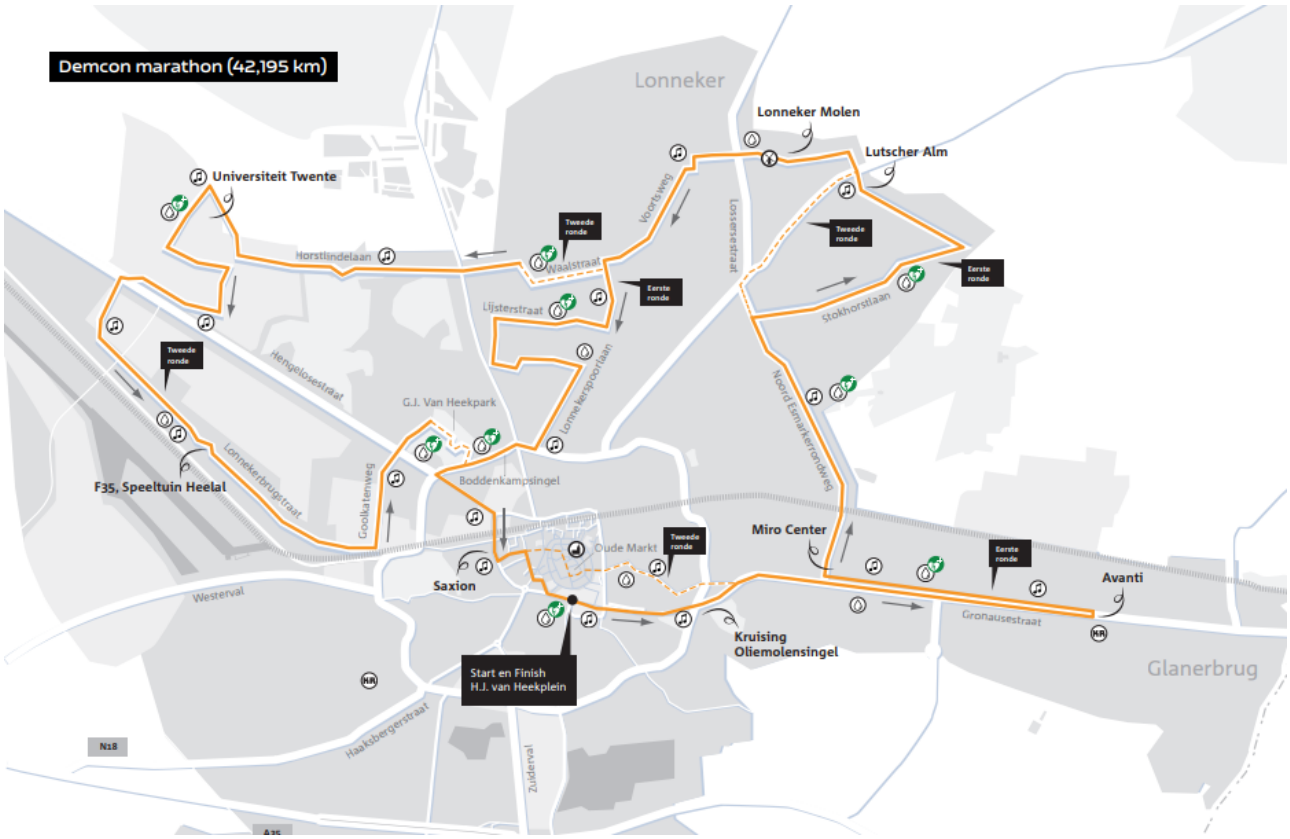

FIGURE 2: MAP OF THE ENSCHEDE MARATHON

2 Literature study

In this literature study, the aim is to review the current state of research on the topics of water stations for endurance activities, features to predict the frequency of water stations, data-driven methods, and data processing. In these topics, research questions 3 to 7 will be answered.

2.1 Features that could impact water consumption

Water consumption at a marathon can be influenced by a variety of factors (Shapiro et al., 1982). Shapiro et al. (1982) state that these include the temperature and humidity of the race day, the speed at which the runner is traveling, and the runner's physiological characteristics.

Temperature and humidity, and the rate at which the surrounding air courses over the athlete's body can impact water consumption because they can affect the rate at which the body sweats and the amount of water that is lost through perspiration (Noakes, 2003). In warmer and more humid conditions, the body will sweat more to cool itself down, leading to increased water loss (Maughan, 2003). On the other hand, cooler and drier conditions will result in less sweating and lower water loss (Maughan, 2003).

Individual physiological characteristics, such as the size and weight of the runner and their sweat rate, can also impact water consumption (Noakes, 2003). Larger runners or those with a high sweat rate may need to consume more water to replace the fluids lost through perspiration (Dancaster & Whereat, 1971).

Finally, the speed at which the runner is traveling can also affect water consumption, as faster runners may sweat more due to the increased energy expenditure (Shapiro et al., 1982). Speed can in turn be influenced by other features.

Vihma (2010) found that in all categories of runners, the air temperature was the single weather parameter with the highest correlation with running speed. In that study, statistically significant correlations with running speed were found in solar shortwave radiation, air relative humidity, and rain.

Knechtle et al. (2019) found a relationship between the running speed and the combination of the wind direction and speed. Next to the environmental factors, the slope of the course also affects the running speed (Margaria et al., 1963). They state that at a specified velocity, running uphill requires a greater rate of metabolic energy supply.

Age is a significant predictor of running speed in marathon runners, with older runners generally having slower running speeds than younger runners (Lara et al., 2014). Gender also plays a role, with men generally having faster running speeds than women (Lara et al., 2014). Body Mass Index (BMI), which is a measure of body fat based on height and weight, has also been found to be significantly correlated with running speed in marathon runners, with lower BMI values generally associated with faster running speeds (Sedeaud et al., 2014).

2.2 Data collected by marathon organizations

Data collection by marathon organizations is an important aspect of event planning and management. Several types of data are typically collected by marathon organizations, including personal information, race results, and feedback from participants.

Personal information is typically collected from participants to register them for the event and to provide them with necessary information about the race. This may include name, age, gender, contact information, and emergency contact information.

Race results are also commonly collected by marathon organizations. This may include information about the time it took for each participant to complete the race, as well as split times on various locations of the route. This data is often used to determine the winners of the race and to provide participants with information about their performance.

Feedback from participants is often collected by marathon organizations. This may be in the form of surveys or interviews and may cover a range of topics including the overall organization of the event, the course, and the facilities. This data is used to make improvements to future events and to gauge the satisfaction of participants.

2.3 Data-Driven Modeling

Data-driven modeling is a modeling approach that relies on collecting and analyzing data to make informed decisions (Hastie, Tibshirani, Friedman, et al., 2009). They state that it is based on the idea that data contains valuable insights that can be used to understand and predict real-world phenomena.

There are several types of data-driven modeling methods, including statistical modeling, machine learning, and econometric modeling (Hastie, Tibshirani, Friedman, et al., 2009). They describe statistical modeling as a statistical technique to fit a model to data and make predictions. Hastie, Tibshirani, Friedman, et al. (2009) define machine learning as a subset of artificial intelligence and statistical techniques that involves the development of algorithms that can learn from data and improve their performance over time. According to Hastie, Tibshirani, Friedman, et al. (2009), machine learning algorithms are commonly used in data-driven modeling as they can analyze large datasets and identify patterns and trends that may not be visible to humans.

Hastie, Tibshirani, Friedman, et al. (2009) point out several advantages and limitations of data-driven modeling. According to them, one of the main advantages of data-driven modeling is that it allows organizations to analyze and understand complex problems by leveraging large amounts of data. It is also able to learn from data and improve performance over time, which can be beneficial for tasks that require a high level of accuracy. One limitation is that the quality of the results is dependent on the quality of the data. If the data is noisy or biased, the results may not be accurate. Additionally, data-driven modeling can be computationally intensive.

2.2.1 Machine Learning

In the field of machine learning, numerous methods can be utilized to analyze data and make predictions (Burkov, 2019). They state that these methods can be broadly classified into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning. Burkov (2019) describes these categories as follows. Supervised learning involves the use of labeled data, which refers to data that has been specifically marked or labeled with the correct output or class. For example, in a study on predicting the likelihood of a person developing a particular disease, the input data (e.g., age, gender, medical history) might be paired with a label indicating whether or not the person ultimately developed the disease. The goal of a supervised learning algorithm is to use this labeled data to produce a model that can take in a feature vector as input and output a label or prediction for that vector. Unsupervised learning, on the other hand, involves the use of unlabeled data, which does not have a predefined output or class. Semi-supervised learning involves the use of both labeled and unlabeled data, while reinforcement learning is a type of machine learning specifically designed to solve sequential problems with long-term goals.

For this study, the available dataset consists of labeled data, as the input data is paired with the output data, the pace of the runner. As such, the scope of the study is limited to supervised machine learning methods. In addition, the goal of this study is to predict continuous values, making supervised regression methods the most relevant choice. This section will provide a detailed description of several relevant supervised machine-learning regression methods. As the focus of this study is on a one-time decision-making problem, reinforcement learning is not applicable (Nasteski, 2017). They state that the use of labeled training data typically leads to more successful results compared to unsupervised techniques, making supervised methods a suitable choice for this study.

Linear Regression

Linear regression is a statistical method that is used to model the linear relationship between a scalar response (also known as the dependent variable) and one or more explanatory variables (also known as the independent variables) (Burkov, 2019). This relationship is visualized by a line in one-dimensional space or a plane in multi-dimensional space, as shown in Figure 3 (Burkov, 2019). Essentially, linear regression seeks to fit a model of the form: $y = wx + b$

where y is the scalar response, x is the explanatory variable or vector of explanatory variables, w is a D-dimensional vector of parameters, and b is a real number. Linear regression is a widely used method for predicting continuous values, and it has the advantage of being relatively simple to implement and interpret (Hastie, Tibshirani, & Friedman, 2009).

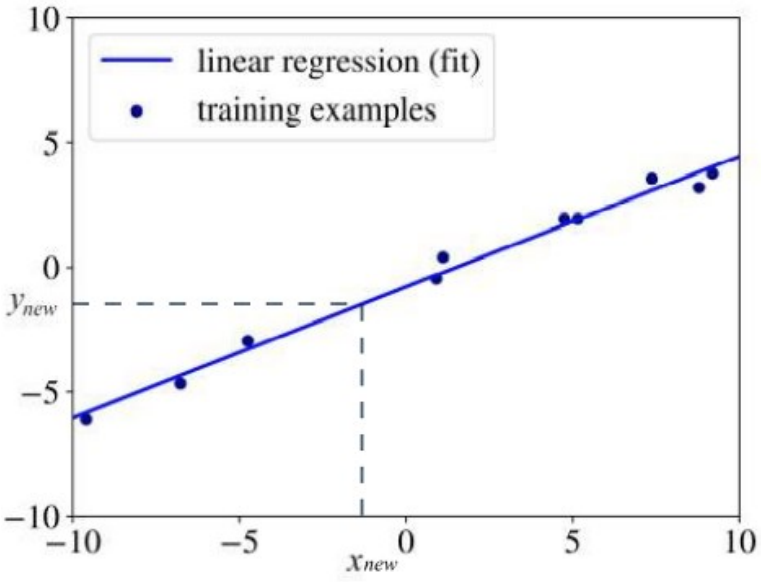

FIGURE 3: LINEAR REGRESSION FROM "THE HUNDRED-PAGE MACHINE LEARNING BOOK" BY BURKOV, 2019

Decision Tree

A decision tree is a type of machine-learning algorithm that can be used to make predictions or decisions based on a set of input data (Burkov, 2019). It consists of a series of branching nodes, each of which tests the value of a specific feature of the input data. Based on the result of this test, the decision tree directs the input data down either the left or right branch of the tree. This process continues until the input data reaches a leaf node, at which point a decision or prediction can be made. Decision trees are often visualized as a tree-like structures with the branching nodes representing the decisions being made and the leaf nodes representing the outcomes or predictions. As shown in Figure 4 by Patel and Prajapati (2018), decision trees can be used to make both categorical and numerical predictions.

Decision trees have several advantages as a machine learning method. They are simple to understand and interpret, and they can handle both numerical and categorical data (Breiman et al., 2017). According to them, they are also relatively easy to implement and can handle large datasets efficiently. They state however that decision trees can also be prone to overfitting, particularly when the tree becomes too deep or the number of features is too high. As such, it is important to carefully tune the parameters of a decision tree model to avoid overfitting and ensure good generalization performance.

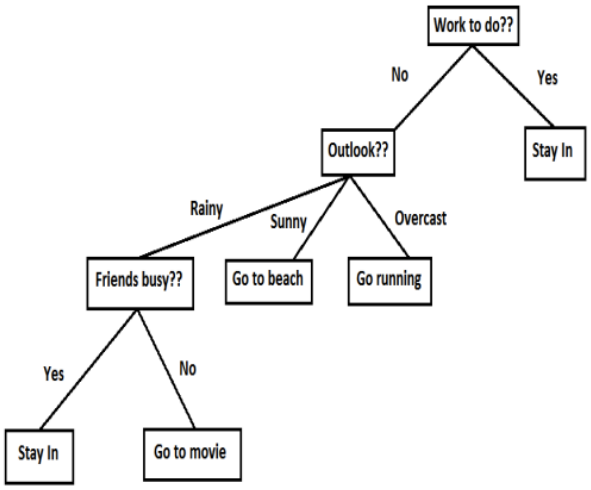

FIGURE 4: DECISION TREE. FROM "STUDY AND ANALYSIS OF DECISION TREE BASED CLASSIFICATION ALGORITHMS" BY PATEL, 2018

Random Forest

Ensemble learning is a type of machine learning that involves training a large number of relatively low-accuracy models and combining their predictions to make a more accurate overall prediction (Zhou, 2012). One of the most popular ensemble learning algorithms is the random forest, which was introduced by Breiman (2001).

The basic idea behind the random forest algorithm is to create many slightly different copies of the training data and use each copy to train a separate decision tree. During the training process, each tree is constructed using a random subset of the features, rather than considering all of the features at each split. This is done to avoid the correlation of the separate trees and ensure that each tree is as independent as possible. If certain features are more impactful, they may appear in the decision nodes of many trees, leading to correlated trees.

The number of trees in the forest and the size of the random subset of features are the most important hyperparameters in random forest learning. The number of trees can be increased to improve the overall accuracy of the model, but this will also increase the computational complexity and may lead to overfitting. The size of the feature subset can also be adjusted to control the complexity of the individual decision trees and prevent overfitting (Breiman, 2001).

Gradient Boosted Trees

Gradient boosting is a type of machine learning algorithm that involves training a series of decision trees to make predictions or decisions based on a set of input data (Friedman, 2002). It is considered one of the most powerful machine learning algorithms because it is capable of creating highly accurate models and handling large datasets (Burkov, 2019). However, in this article, it is stated that it can also be slower to train than some other algorithms, such as random forests.

The basic idea behind gradient boosting is to sequentially add decision trees to the model to improve its performance. The algorithm begins by training a simple decision tree, and then the performance of the tree is measured using a loss function. Other trees are then added to the model

in an attempt to lower the loss until a local optimum is reached. The trees added to the model can be as simple as a single decision, or "stump."

Gradient boosting has several advantages as a machine learning method. It is highly effective at reducing bias and variance, and it can handle a wide range of data types and distributions (Friedman, 2002). They also point out that it is relatively simple to implement and can be easily parallelized to speed up training. However, it can also be sensitive to the choice of hyperparameters and can be prone to overfitting if not properly tuned (Friedman, 2002).

Artificial Neural Network

Artificial neural networks (ANNs) are a type of machine learning algorithm that is inspired by the structure and function of biological neurons (Haykin, 2009). They are characterized by their deep learning architecture, which means that most of the model parameters are learned from the outputs of preceding layers rather than directly from the input data (Burkov, 2019). ANNs consist of layers of interconnected nodes, each of which represents a vector function that takes one or more input values and produces a single output value. The nodes in the first layer receive the external data as input, and the nodes in the final layer produce the final output values. Wang (2003) visualized this in Figure 5.

ANNs are structured in a similar way to biological neural networks, with the nodes in each layer only corresponding to nodes in the layers immediately preceding or following it. During the training of a model, the parameters of the vector functions are optimized using a particular cost function to improve the accuracy of the model. ANNs are capable of learning complex relationships in data and can handle a wide range of data types and distributions (Haykin, 2009). However, they can also be sensitive to the choice of hyperparameters and can be prone to overfitting if not properly tuned (Haykin, 2009).

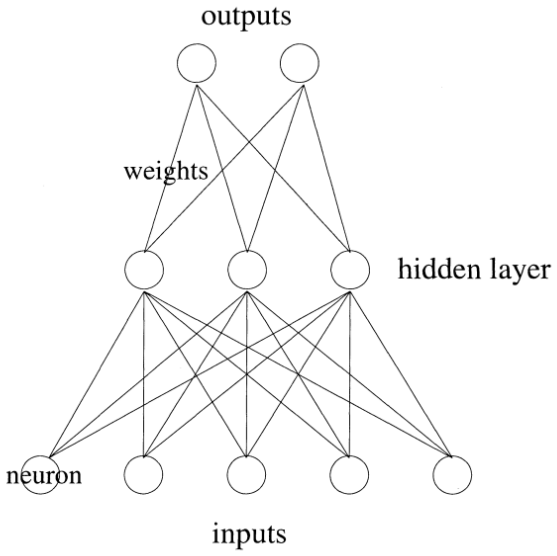

FIGURE 5: ARCHITECTURE OF NEURAL NETWORK FROM "ARTIFICIAL NEURAL NETWORK" BY WANG, 2003

2.3 Data processing

The process of developing a machine learning algorithm involves several key steps, including data cleaning and outlier removal, feature selection (if the number of features is large), and hyperparameter fine-tuning (Vabalas et al., 2019). Cross-validation is often used to control overfitting during this process. In the following section, each of these steps is further elaborated on to understand their role in the machine learning process.

2.3.1 Feature selection

Feature selection is a crucial step in the machine learning process, as it allows for a reduction in computation time, improved prediction performance, and a deeper understanding of the data (Chandrashekar & Sahin, 2014). Chandrashekar and Sahin (2014) identify three main types of feature selection methods: filter, wrapper, and embedded.

Filter methods use statistical techniques to evaluate the impact of each input variable on the output variable, and filter out the least relevant variables. These methods are popular for their simplicity and practical applications.

Wrapper methods, on the other hand, involve the creation of multiple models using different subsets of input variables, and the selection of the variables used in the most successful model. There are many wrapper feature selection methods available.

Embedded methods, such as tree algorithms like Random Forest, integrate feature selection directly into the machine learning algorithm. These methods are well-known and widely used.

2.3.2 Validation methods

The process of machine learning typically involves three main steps: training, validation, and testing (Raschka, 2018; Reitermanova, 2010). These steps are designed to prevent overfitting by using different segments of data for each step (Reitermanova, 2010).

First, the training data is fed into a learning algorithm to generate one or more models, which consist of a specific machine-learning technique with a set of hyperparameters (Raschka, 2018). Next, these models are validated to select the best-performing model. Finally, the chosen model is tested on the testing data set to assess its performance. This final step is important because the results from the validation step may be biased due to the selection of the best-performing model (Raschka, 2018).

Alternatively, cross-validation can be used to iteratively develop multiple models on different portions of the data, rather than training a single fixed model as in a train/test split (Vabalas et al., 2019). One example of cross-validation is K-Fold cross-validation, which allows for the use of all the data for training and validation. However, using the same data for both development and validation can result in over-optimistic performance estimates (Varma & Simon, 2006). To address this issue, Krstajic et al. (2014) propose nested cross-validation as a solution that avoids this problem while still being efficient with the data.

Nested cross-validation consists of an outer loop and an inner loop, shown in the figure (*Nested Cross Validation*, 2020). The outer loop is repeated n times, generating n unique test sets. The inner loop is used to select the best model by training and validating different models on the train set of the outer loop. The inner loop is repeated m times, generating m unique validation sets. In total, nested cross-validation will involve n x m trained models (Krstajic et al., 2014). This method provides almost unbiased performance estimates (Varma & Simon, 2006).

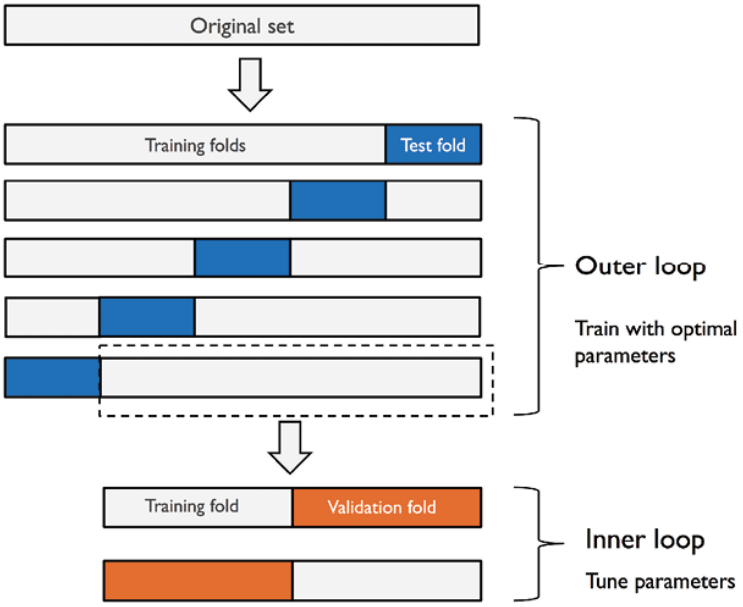

FIGURE 6: NESTED CROSS VALIDATION FROM "VITALFLUX.COM" BY KUMAR, 2020

2.3.3 Hyperparameters tuning

Hyperparameter tuning is an important step in the machine learning process, as it allows for the optimization of the parameters that influence the performance of the algorithms. There are various methods available for finding the best sets of hyperparameters to test, including full factorial design (Montgomery, 2017). This method involves choosing a set of values for each hyperparameter and using grid search to test a set of combinations. However, the number of combinations can become unwieldy when there are many hyperparameters.

An alternative method is a random search, proposed by Bergstra and Bengio (2012), which generates a random set of combinations to test. According to Bergstra and Bengio (2012), random search is less influenced by an increase in the number of hyperparameters and performs better than full factorial design when certain hyperparameters are more important than others.

Other methods for hyperparameter tuning include Bayesian optimization, which uses Bayesian principles to model the function being optimized and search for the optimal set of hyperparameters (Snoek et al., 2012), and genetic algorithms, which use principles of natural evolution to search for the optimal set of hyperparameters (Goldberg et al., 1989).

Several studies give some indications of the optimal ranges for the hyperparameters of a certain algorithm. For linear regression, it is recommended to use a range of values for the regularization strength (C) such as 0.01 to 1, and a range of values for the L1 ratio (Lasso and Elastic Net

regularization) such as 0 to 1 (Boyd et al., 2011). For decision trees, common ranges to test include 1 to 20 for the maximum depth and 2 to 100 for the minimum samples per leaf (Hastie et al., 2009). For random forests, the number of estimators can be tested in the range of 10 to 1000, and the maximum depth can be tested in the range of 1 to 20 (Breiman, 2001). For gradient-boosted trees, the learning rate can be tested in the range of 0.01 to 1, and the number of estimators can be tested in the range of 10 to 1000 (Friedman, 2002). For artificial neural networks, the number of hidden layers can be tested in the range of 1 to 5, the number of neurons per layer can be tested in the range of 10 to 100, the learning rate can be tested between 0.001 and 0.1 and the number of hidden units can be tested between 50 and 500 (Goodfellow et al., 2016).

It is important to note that all studies stressed that these ranges are only recommendations and may not be optimal for all datasets and problems. Careful experimentation and testing are necessary to determine the optimal ranges for the hyperparameters of each machine-learning algorithm.

2.4 Summary

Water consumption during a marathon can be affected by various factors, including temperature, humidity, running speed, and course conditions. Marathon organizations often collect data, such as participant information and feedback, to improve event planning and management. This data can be used through data-driven modeling, which involves collecting and analyzing data to make informed decisions. Machine learning is a type of data-driven modeling that involves the development of algorithms that can learn from and improve with data. There are four categories of machine learning: supervised, unsupervised, semi-supervised, and reinforcement learning. For this study, the dataset consists of labeled data and the goal is to predict continuous values, making supervised regression methods the most suitable choice. Linear regression is a statistical method used to model the linear relationship between a dependent and independent variable, while decision trees involve branching nodes that test specific features of input data. Random forests are an ensemble learning algorithm that trains multiple decision trees on different copies of the training data, and gradient boosting involves training a series of decision trees using gradient descent optimization. Artificial neural networks are inspired by the human brain and consist of interconnected nodes, and support vector machines are a supervised learning algorithm that creates a decision boundary to separate data. Hyperparameter tuning is the process of optimizing the parameters that influence the performance of machine learning algorithms, and cross-validation is used to prevent overfitting and assess the performance of the chosen model.

3 Methodology

The research question for this study was: " *How can Enschede Marathon use a data-driven method to find the optimal placement frequency for water stations to achieve a sufficient performance of runners?*" The purpose of the study was to find a solution to the problem of how to use data to understand the impact of external factors on the placement of water stations at the Enschede Marathon. The study aimed to identify the current situation at the marathon, including the current water station frequency method and the features taken into consideration, and to review the literature to create a framework of findings on water management at endurance sports events and the implementation of data-driven methods.

The study also investigated the performance of different models and the impact of different features on the prediction of water station locations and explored the implications of the results for the Enschede Marathon. The methods used for this part of the research are described in this chapter. The methodology consists of the collection of data, the processing of this data, and the machine learning techniques to arrive at the findings.

3.1 Data Collection

In this study, the initial data collection effort involved attempting to obtain raw data from a selection of marathons. However, efforts to obtain such data directly from multiple marathon organizations were unsuccessful. As a result, a web scraper was developed to extract runner data from marathons that had made this data available on the "MyLabs Sporthive" platform. The specific marathons included in this study were the TCS Amsterdam Marathons of 2018, 2019, and 2021, the NN Marathons of Rotterdam in 2019, 2021, and 2022, and the EDP Maratona de Lisboa in 2021. Figure 29 in Appendix C shows how the data from the Amsterdam Marathon was scraped and stored in an Excel file.

The data for each runner from these marathons included the marathon name, gender, age category, and multiple split times, along with the locations of these split times. Each of these data points was treated as a separate data entry, comprising the marathon name, runner gender, age category, split time, and location. Using GPS files of the routes of each marathon and the locations of the split times, additional data points were also added to the individual data entries, including the direction of the course relative to the current location, as well as elevation changes (both up and down). The total number of water stations passed by each runner was also included in the data based on information that was provided by the marathon, either through correspondence or on the website. The included data was based on the availability of the data and insights gained in the literature.

In addition, weather data from the day of each race was collected from Wunderweather, a service that provides historical, half-hourly weather updates for specific locations. The weather data included temperature, humidity, wind speed, and wind direction. This data was incorporated into the data entries for each runner. Both the starting time of each runner and the pace of that runner were taken into account to use the correct time and place to find the corresponding weather values.

3.2 Data processing

Before the data could be used for machine learning algorithms, it needed to be processed. The first step in this process was to perform a general analysis of the available data to gain a better understanding of it. This analysis focused on identifying any patterns that might exist between the input features and the output feature (pace) with the use of scatter plots. These scatter plots were later on also used to provide context for certain predictions and to understand where these predictions were coming from.

Next, the data was cleaned by removing missing data points and outliers. Data entries with missing information were removed from the dataset, and outliers were identified based on a minimum and maximum pace (min/km) determined by the fastest times for the marathon and a calculation that included three times the standard deviation of the pace added to the mean pace.

Finally, some feature selection analyses were conducted to possibly exclude certain features. Correlation statistics were used to determine the level of correlation between different features, and highly correlated features were considered for removal as they would have similar predictive power. An embedded tree-based feature importance method was also used to rank the importance of each feature based on a tree-based machine learning algorithm.

3.3 Machine learning

In this study, machine learning algorithms were applied to process data to make predictions or classify data points. In the validation step, multiple algorithms were compared and their parameters were optimized. The chosen validation method was nested cross-validation, which is a robust method for evaluating the performance of machine learning models as was shown in the literature study. Nested cross-validation involves splitting the data into an inner loop and an outer loop, and using the inner loop for model selection and the outer loop for model evaluation. The inner and outer loops in this study were both split into three sets, based on a tradeoff between computation time and the quality of the model, as well as some standard values used in the literature. The programming of the algorithms and the nested cross-validation procedure were carried out in Python using the Skikit learn module. An example of this nested cross-validation is shown in Figure 30 in Appendix C.

The quality of the different algorithms was measured using the Mean Square Error (MSE) metric. MSE is a commonly used statistical metric that measures the average squared difference between the predicted values and the actual values in a dataset. The results of the nested cross-validation procedure included three outcomes for each algorithm, along with a set of optimized hyperparameters. The best results of each algorithm were compared and the algorithm with the highest performance and the best set of hyperparameters was selected for further analysis.

In the analysis of results, the chosen machine learning algorithm was applied to a test dataset to evaluate the impact of individual features on the target variable (pace). To isolate the effect of each feature, other features were held constant at their average values observed during the Enschede Marathon of 2022. For example, when testing the effect of humidity on pace, the temperature was set to the average temperature during that marathon.

In addition to testing individual features, combinations of features were also evaluated. However, it was not feasible to test all possible combinations due to the large number of options. Instead,

combinations that were expected to yield interesting results based on experience or literature were selected for testing.

4 Results

In this chapter, the results of the data processing, performance, and findings of the developed machine learning models, and the analysis of the results are shown.

4.1 Data processing

Before the machine learning could be performed, the data needed to be pre-processed. This consisted of data cleaning and feature selection. The data cleaning step was crucial to ensure that the data used was accurate and relevant. Figure 7 was created to get a better understanding of how the data should be cleaned. This figure showed how the data was distributed and pointed to possible outliers. Based on the information in the figure, outliers were removed when the pace of a data point exceeded the point of three times the standard deviation above the average. On the other hand, lower paces were removed when the pace of a certain data entry was lower than the number one runner of that marathon with a margin of 10 percent. Additionally, missing data points, in this case only the pace, were removed from the data set.

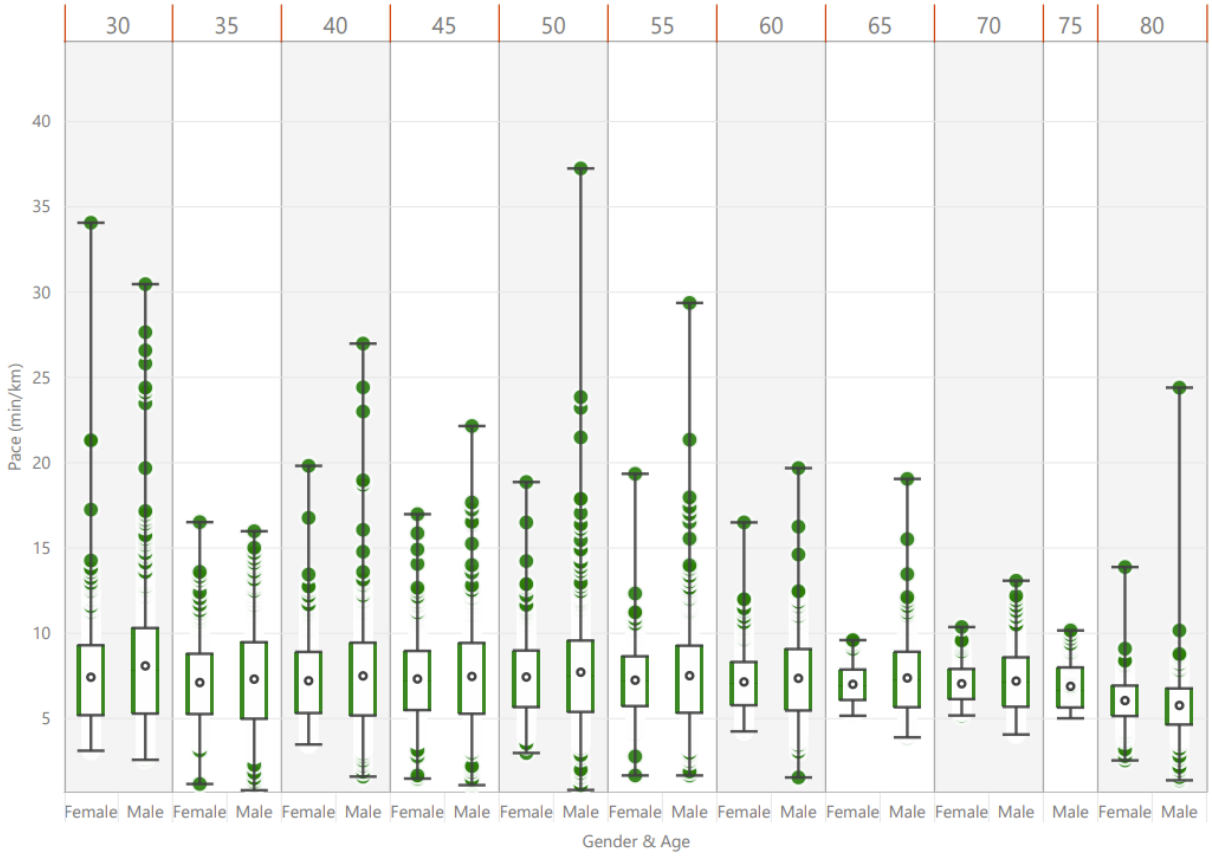

FIGURE 7: BOXPLOTS OF THE PACE FOR GENDER & AGE

In the feature selection step, the aim was to improve the speed of the machine learning models while still capturing the relevant information. A correlation diagram, shown in Figure 8, was created to study the correlation between the different features. As stated in the methodology, the included features were based on the availability of the data for each feature and the results from the literature study. From this diagram, it was evident that features like elevation up and down, water stations, and location were closely correlated. Another feature selection method was performed to possibly remove features that didn't significantly impact the outcome of the machine learning models. The results of this method are shown in Figure 9 and indicated that Humidity, Windspeed, and Temperature seemed to have significantly more impact on the pace than the other features. However, these features were kept in the data set for the machine learning models as the goal of this research was to gain insight into the effect of several features on the pace of the runners.

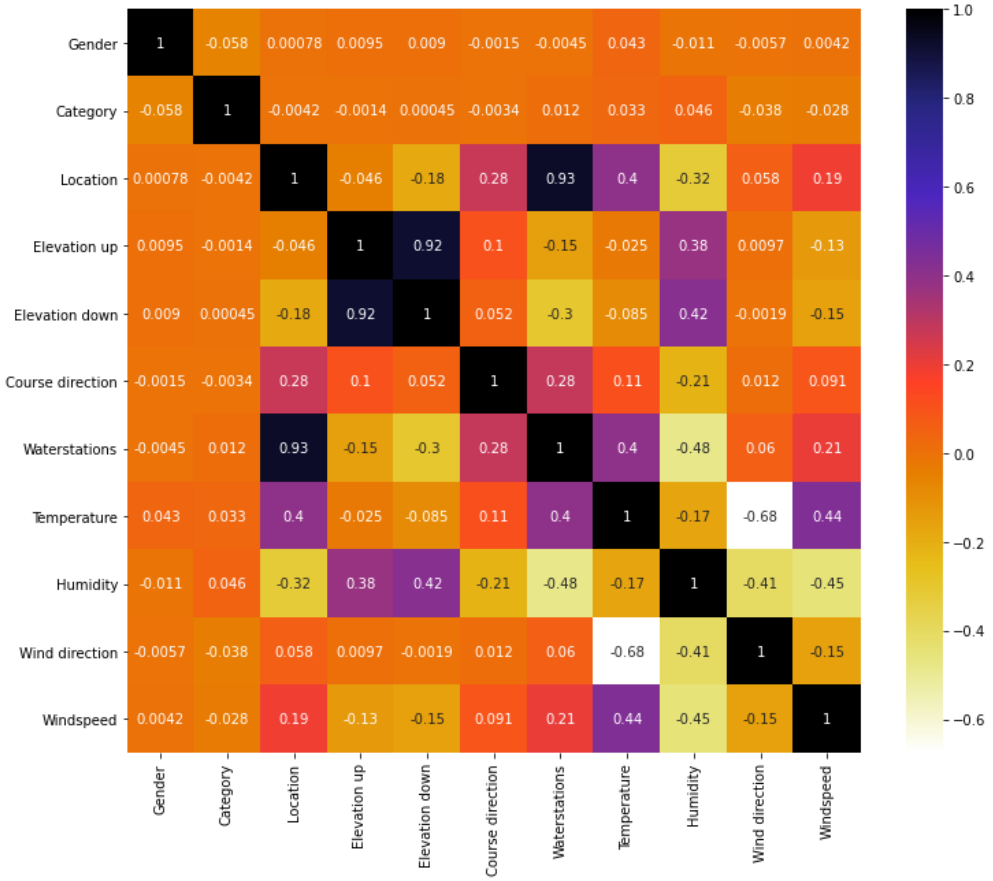

FIGURE 8: CORRELATION BETWEEN FEATURES

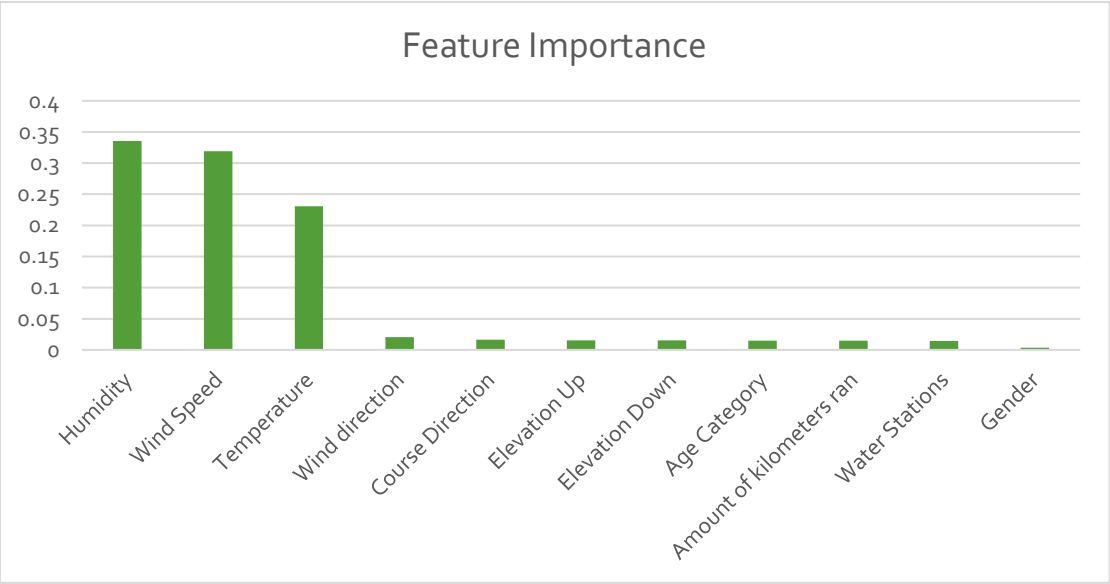

FIGURE 9: FEATURE IMPORTANCE

4.2 Creating a machine learning model

The next step in the research process was to apply machine learning algorithms to the processed data to evaluate their performance in predicting the pace of runners in a marathon. To compare the performance of the different algorithms and to tune the hyperparameters of each algorithm, nested cross-validation was used.

The nested cross-validation consisted of three outer loops and five inner loops. The choice of three outer loops and five inner loops was based on a combination of time restrictions and values found in the literature. The time restriction was based on the time it took for one iteration of each algorithm.

Figure 10 shows the time it took for one iteration of each algorithm. As can be seen from the figure, Decision Tree and Linear Regression took close to no time, while Neural Network took much more time, even though the option for multiple layers was removed to reduce the time required. The other algorithms, Gradient Boosting Trees and Random Forest were in between these two extremes in terms of the time taken.

The algorithms that were tested were Decision Trees, Gradient Boosting Trees, Linear Regression, Random Forest, and Neural Networks. These algorithms were all tested with a different range of parameters, as shown in Figure 25 in Appendix A, based on the values found in the literature. The ranges of parameters were selected to ensure that each algorithm was evaluated fairly and with a range of possible hyperparameters.

FIGURE 10: TIME OF ONE ITERATION PER ALGORITHM

The nested cross-validation was used to evaluate the performance of each algorithm in terms of MSE. The results of the nested cross-validation are shown in Figure 11A, which displays the MSE for each algorithm. As can be seen from the Figure, linear regression, and neural network did not perform very well, while the other algorithms performed similarly. Figure 11B shows the results of only the other three algorithms, and it can be seen that Gradient Boosting Trees performed best, followed by Random Forest and Decision Tree.

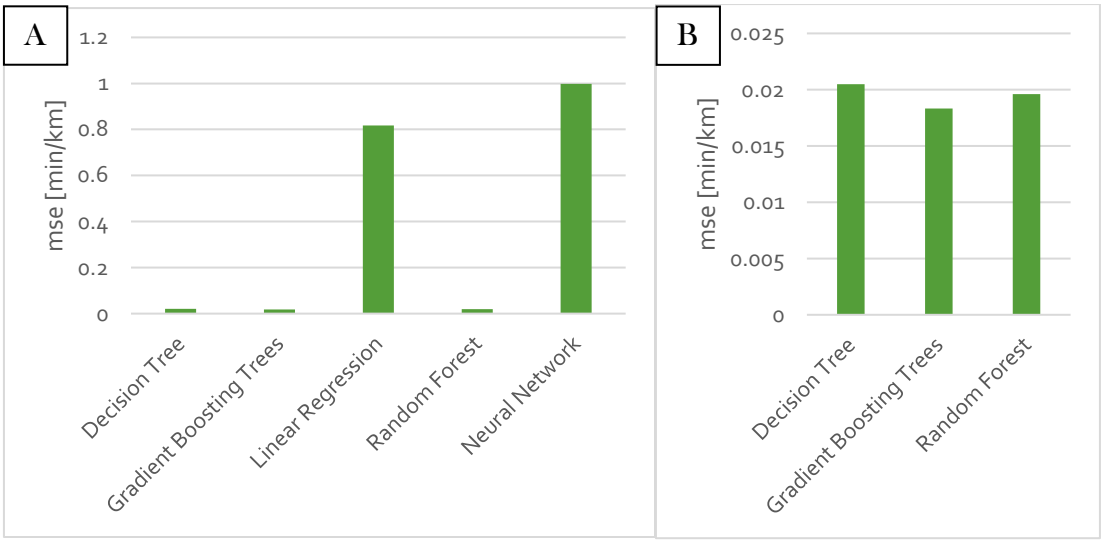

FIGURE 11: MEAN SQUARE ERROR (ME) PER ALGORITHM (A &B)

To further study the performance of the algorithms, a violin plot was created, as shown in Figure 12. This plot displays the distribution of the error for each algorithm, and it can be seen that no algorithm has a bias. As expected, linear regression and the neural network had a lot of errors between -2 and 2, while the other algorithms the had most errors around 0. Gradient Boosting Trees had the least number of extreme outliers.

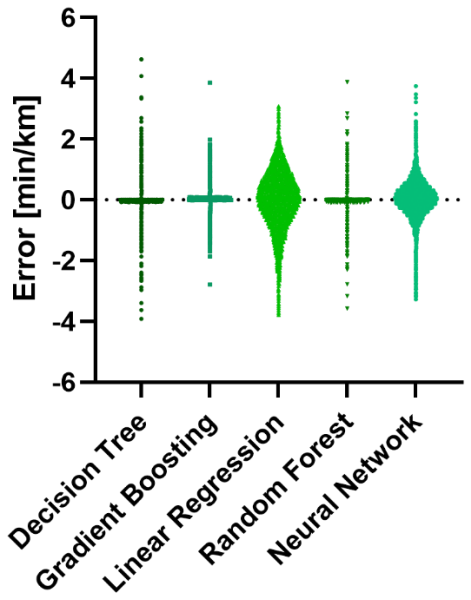

FIGURE 12: VIOLIN PLOT MACHINE LEARNING MODELS

Finally, the learning curve of the machine learning algorithms was analyzed, as shown in Figure 13. The x-axis in Figure 13 represents the number of iterations, while the y-axis represents the MSE score. These scores reflect the performance of the model on the training set and validation set, respectively, and are used to evaluate the model's ability to generalize to new data.

Based on Figure 13, it appears that the model is underfitting the training data as the training score remains low and has only a slight increase. The underfitting is also supported by the violin plot of Figure 12, which shows a large number of predictions with high error. The validation score further indicates that the model is not able to generalize well to new, unseen data, as it exhibits a decreasing trend with significant fluctuations. These observations suggest that the model lacks the necessary complexity to capture the underlying patterns in the data. The learning curves of the other machine learning algorithms are presented in Figures 26, 27, and 28 in Appendix B, and they exhibit similar results, except for Linear Regression, which shows more convergence between the validation score and the training score, but still ends with a larger gap than the other algorithms. To overcome this problem of underfitting, the complexity of the model was attempted to be increased by adding nodes to the tree based models, this did however not have any effect. Increasing the complexity by adding features was also not an option as all features were already included.

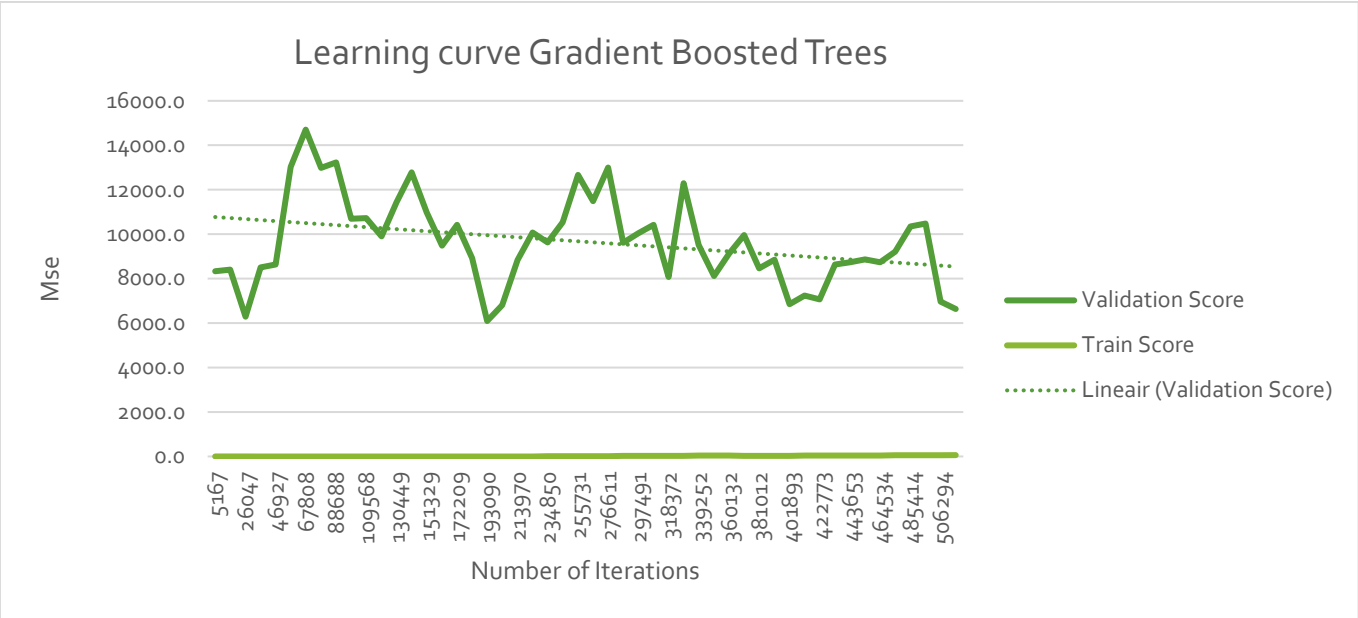

FIGURE 13: LEARNING CURVE GRADIENT BOOSTED TREES

4.3 The predicted effect of features on pace

In this section, the results of the machine learning model are presented, specifically focusing on the best-performing algorithm, Gradient Boosting Trees. The impact of various features and combinations of features on the pace of the runners were evaluated using this algorithm. In the following subsections, the results of this analysis are discussed. Each subsection consists of a Figure with Pace on the Y axis and one feature on the X axis. On the graph, a line shows the predicted values for the pace by the machine learning algorithm based on the changing values of the feature in the x-axis. All other feature values, that were used as input values to the machine learning algorithm, were either based on the Enschede marathon of 2022 like temperature and humidity or chosen randomly like age category and gender. The temperature was 15 degrees Celsius, the humidity was 50% and the wind speed was 6.6 meters per second. All Figures in this section also show the scattered original data to provide some context to the predicted line. The scattered data could indicate where the predicted values come from.

Humidity

The effect of humidity on the pace of runners was analyzed using the trained machine learning algorithm. Figure 14 shows the predicted pace according to this algorithm, along with the scattered input data. Clear patterns were observed in the scattered data in form of lines. Every line in the scatter plot originates from one marathon and one measuring point. Often the humidity steadily increases or decreases over the courses of a certain time period, especially since the weather information is half hourly and the rest of the data between these points is interpolated. The runners at most marathon start at the same time. Slower runners will reach the measuring point later, where they will find a humidity that is just a bit higher or lower than the runner before him or her. This leads to the lines in the scattered data. The predicted line showed a slight decrease in pace as humidity increased.

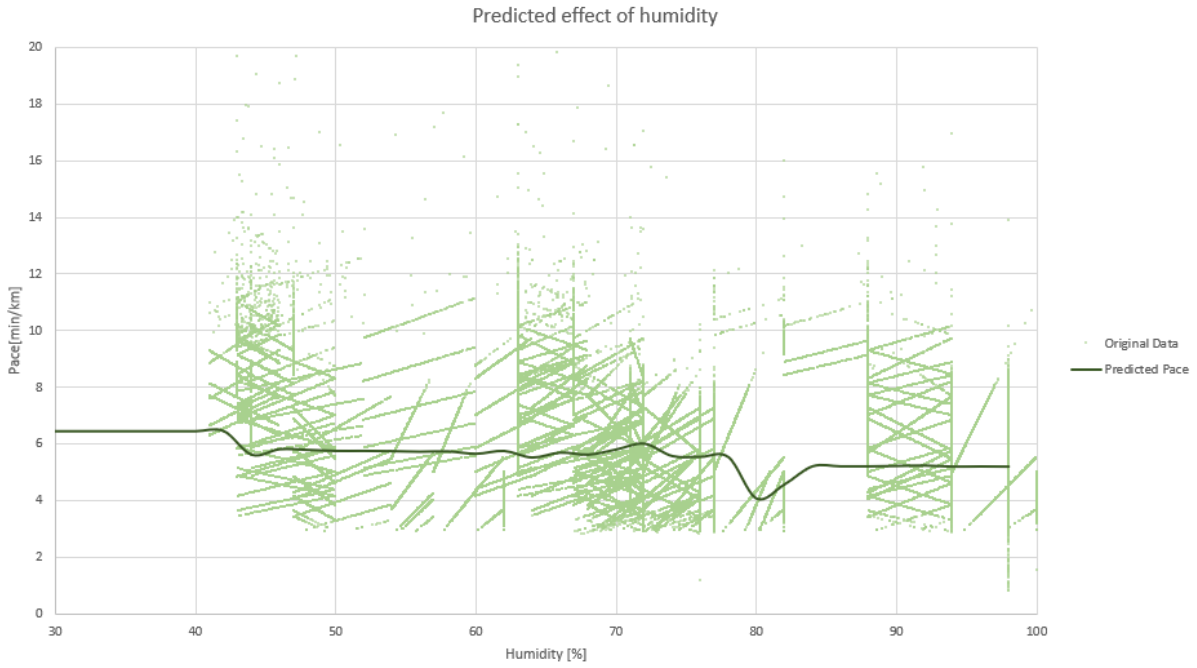

FIGURE 14: PREDICTED EFFECT OF HUMIDITY

Temperature

The effect of temperature on the pace of runners was analyzed using the machine learning model. Figure 15 shows the predicted pace compared to temperature, along with the original scattered data. Similar to Figure 14, the scattered data showed a non-random distribution, suggesting a relationship between temperature and pace. The figure revealed that the pace tends to increase as the temperature increases. The predicted line seems to follow the scattered data and could therefore be explained by the original data.

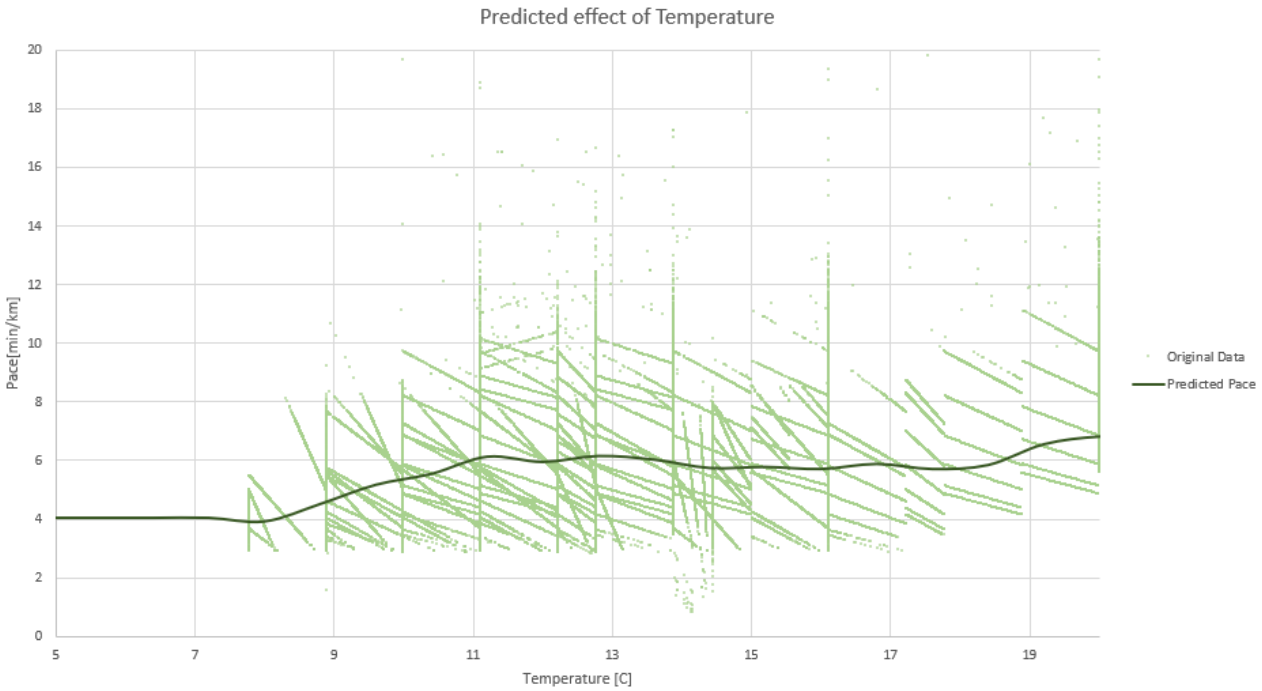

FIGURE 15: PREDICTED EFFECT OF TEMPERATURE

Temperature & Humidity

Figure 16 presents the combined effect of temperature and humidity on pace. This Figure shows the effect of humidity on the pace at four different temperatures. It was found that at lower percentages of humidity, higher temperatures had a greater impact on the pace compared to higher humidity percentages. It appears that the projected lines are positioned lower than what the dispersed data would have indicated, although the scattered data seems more dense at the lower pace.

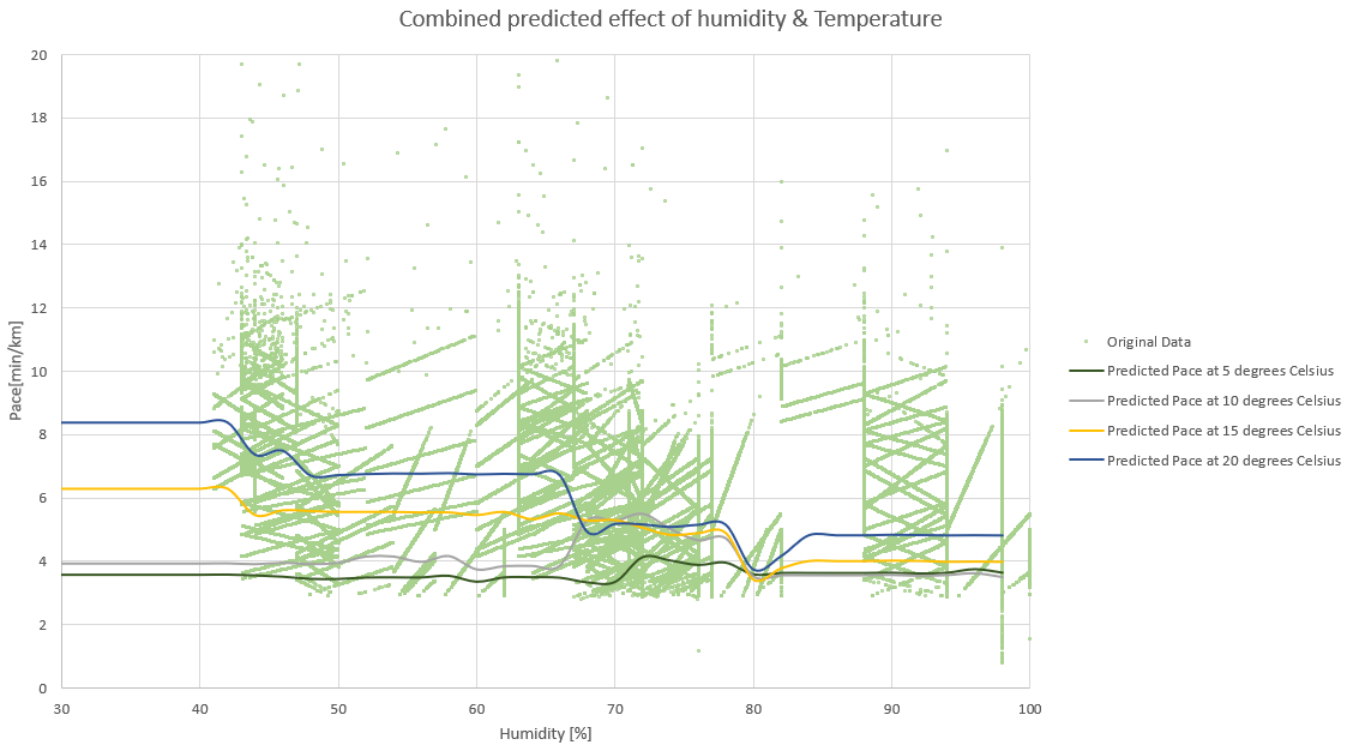

FIGURE 16: EFFECT OF TEMPERATURE AND HUMIDITY

Age Category

The effect of age on pace was also analyzed. Figure 17 shows the predicted effect of age on pace, which indicated that there is only a very slight increase in pace as the runner gets older. At first sight, the predicted line is quite low compared to the original data. The original data is however scattered in lines, which makes that the distribution is not as clear as real scattered data.

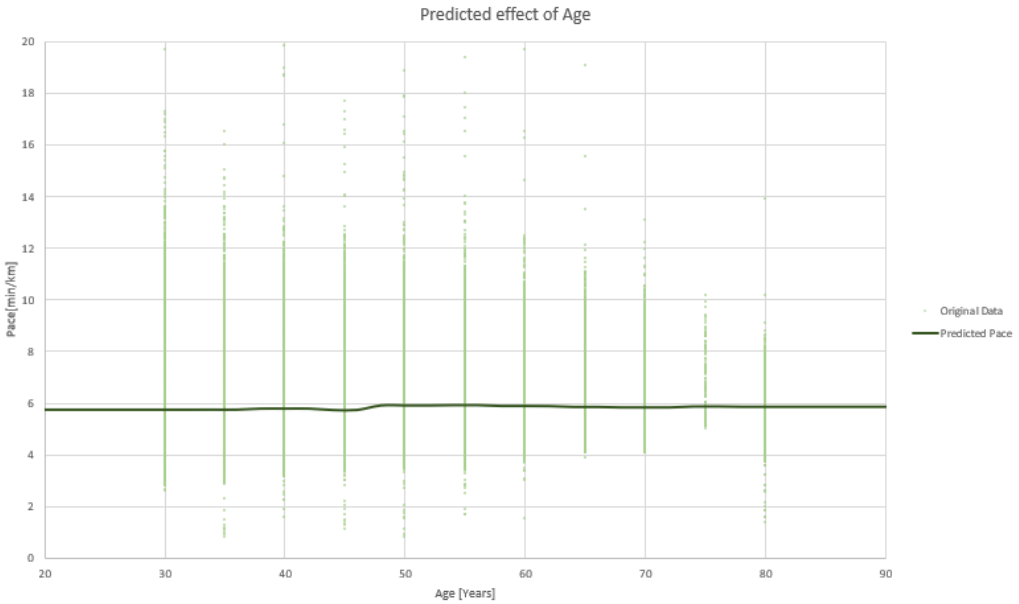

FIGURE 17: PREDICTED EFFECT OF AGE

Water Stations

The effect of water stations on the pace of runners was evaluated. Figure 18 shows the predicted effect of water stations on pace, revealing a substantial decrease in pace between 2 and 10 water stations. Interestingly, this result did not immediately correspond with the scattered original data shown in the Figure. This could however also be explained by the scattered data being on one line, as it was in Figure 17.

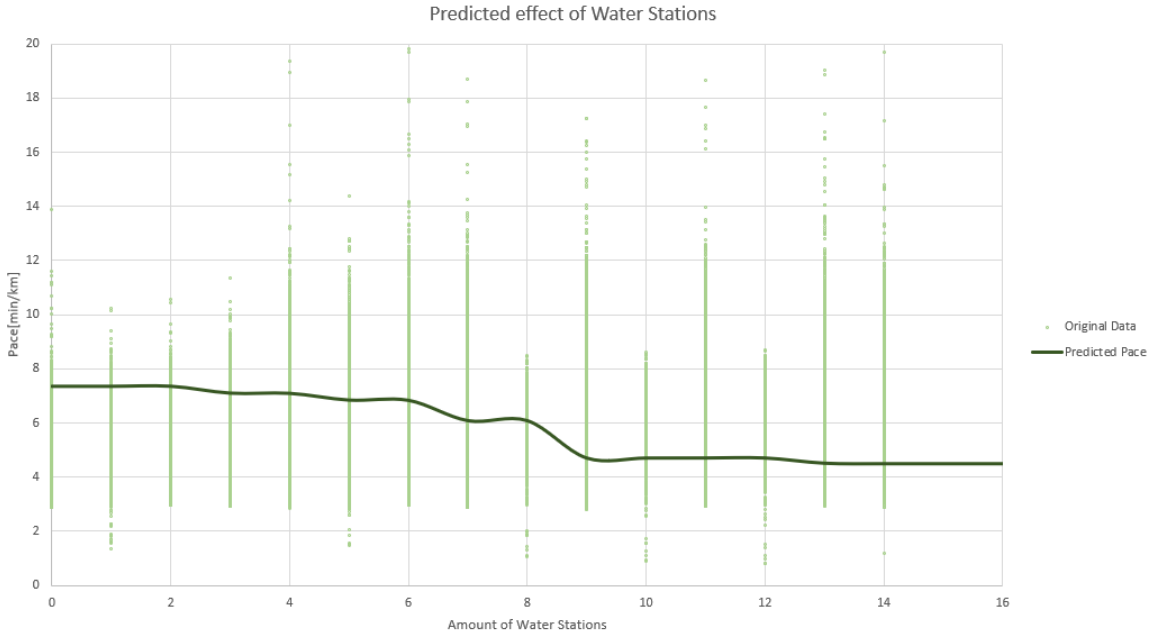

FIGURE 18: PREDICTED EFFECT OF THE WATER STATIONS

Distance

The impact of distance on pace was analyzed. Figure 19 shows the predicted effect of distance on pace, indicating that the pace tends to decrease as the number of kilometers increases. However, this finding seems to be inconsistent with the input data scatterplot also shown in Figure 19. The distribution of the data points is however not clear, as all points are on one line.

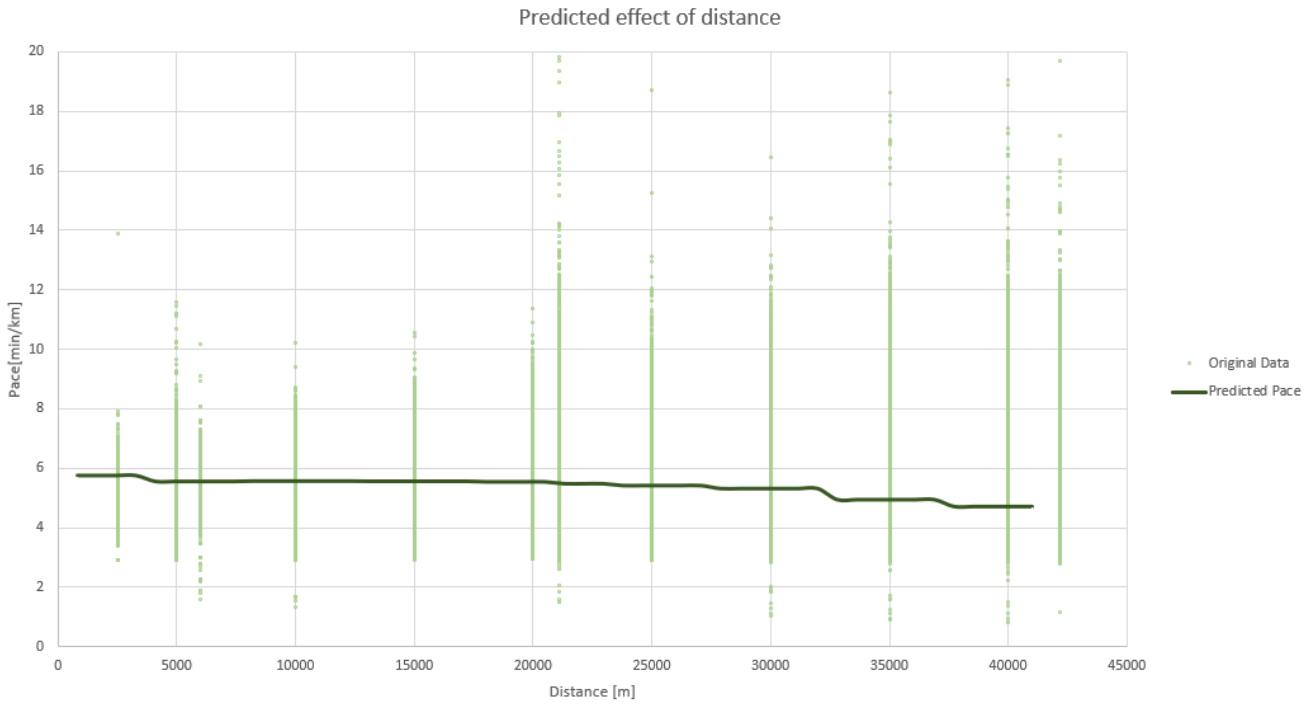

FIGURE 19: PREDICTED EFFECT OF DISTANCE

Water Stations & Distance

Figure 20 shows the predicted effect of water stations at different distances during the marathon. The result looks like a combination of Figures 18 and 19, which show similar trends. Distance does not seem to influence to impact of water stations on pace.

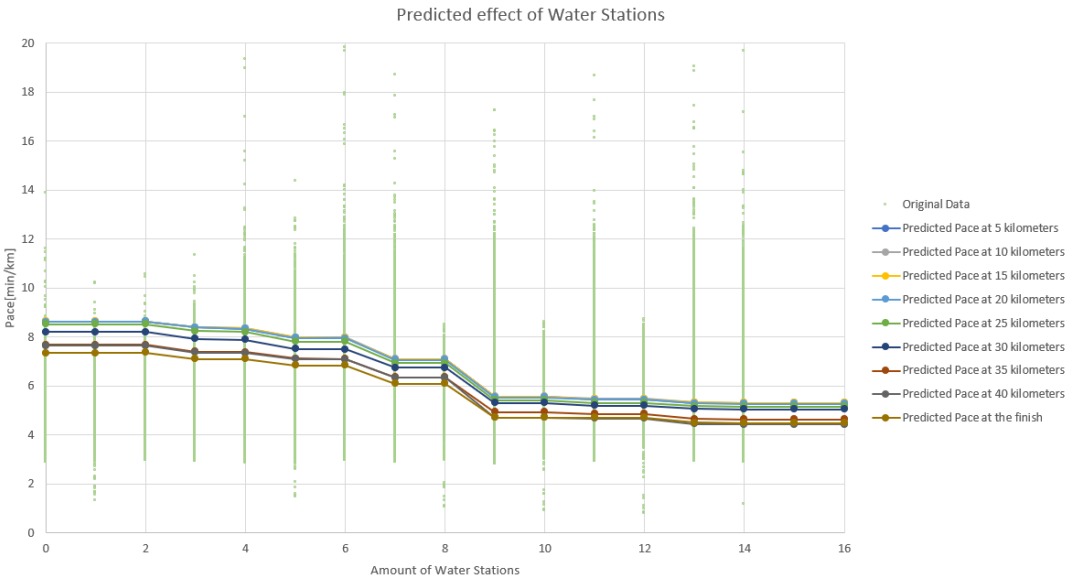

FIGURE 20: PREDICTED EFFECT OF WATER STATIONS AT DIFFERENT DISTANCES

Water Stations & Temperature

Figure 21 shows the combined effect of water stations at different temperatures. It shows that the amount of water stations at 5, 10, and 15 degrees Celsius does not seem to make a big impact, while at 15 degrees, the pace goes down considerably when the number of water stations rises around 9.

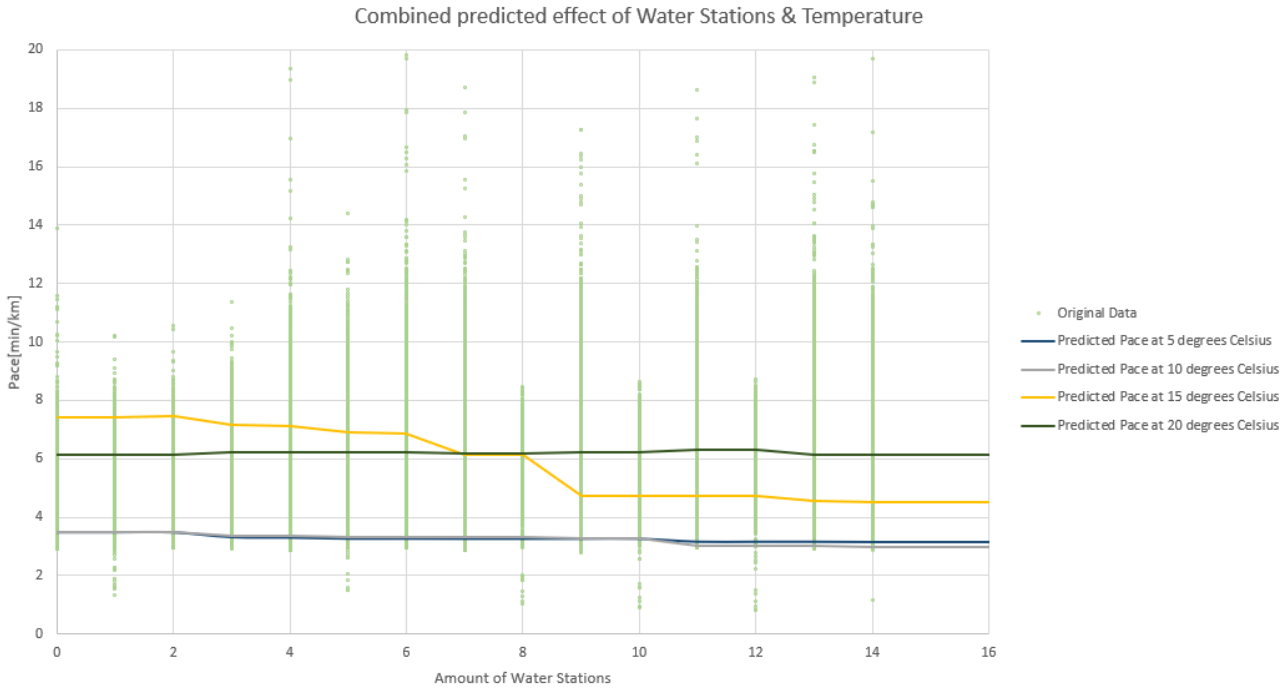

FIGURE 21: THE EFFECT OF WATER STATIONS & TEMPERATURE

Wind

Finally, Figure 22 shows the scattered original data of windspeed on the x-axis and pace on the y-axis. Three predicted pace lines are also shown in the figure. Each line represents a different direction of the wind compared to the running direction. The effect of side, back, and headwinds are shown. The Figure shows no clear difference between headwind and backwind.

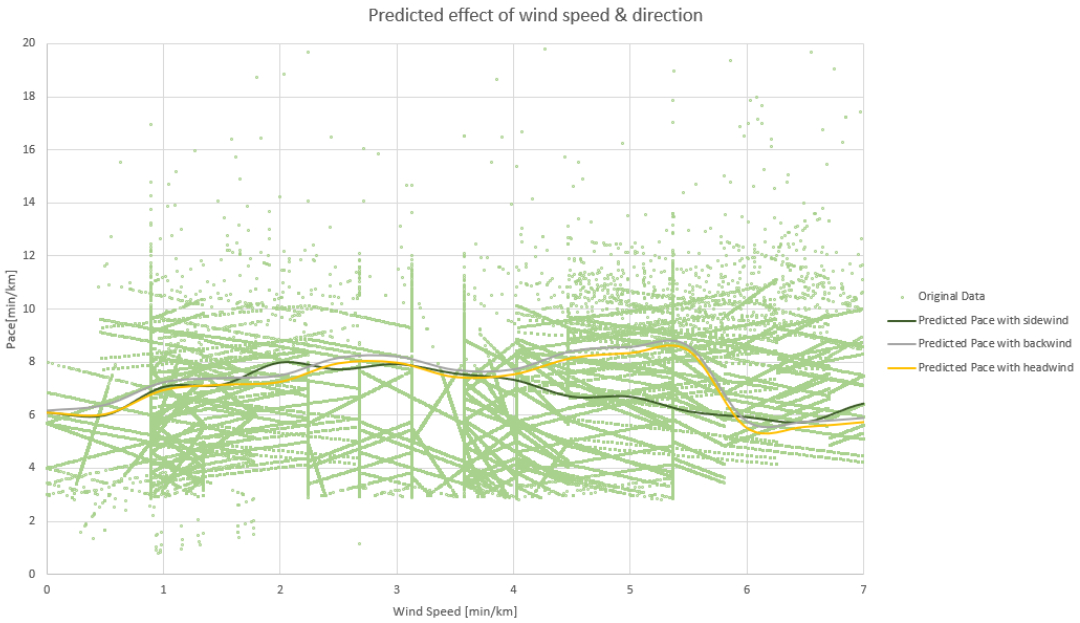

FIGURE 22: PREDICTED EFFECT OF WIND

4.4 A selection of possible scenarios for the Enschede Marathon

To provide some further insights for the Enschede Marathon, several interesting scenarios were developed with different weather conditions. The effects on the predicted water stations were shown for these scenarios. Figure 18 already shows the effect of water stations on pace in the mild weather conditions of the 2022 Enschede marathon. These scenarios were focused on more extreme situations to provide the organization with useful insights. Features like age and gender are random and marathon-specific features are all based on the Enschede Marathon as it was in the previous section.

Scenario 1

The first scenario describes a weather situation for the Enschede Marathon where there is a lot of wind and the temperature and humidity are low. A lot of wind is characterized by a wind speed of 14 meters per second, as this is the windspeed that is classified as strong on the chart of Beaufort. Low temperature is defined as 0 degrees Celsius and low humidity as 0 percent. This scenario should provide insights into situations that with similar values for these weather conditions or values that approach these conditions for the Enschede Marathon.

Scenario 2

The second scenario describes a similar situation to scenario 1. In this scenario however, there is no wind instead of the strong wind of scenario 1. Temperature and humidity are still very low here.

Scenario 3

The third scenario describes a situation which can be seen as the opposite of scenario 2 in terms of weather conditions. The scenario describes a possible future situation where the strong wind has a speed of 14 meters per second. The temperature is 25 degrees Celsius, which can be described as very high for the Netherlands in April. The humidity in this scenario was also set at a high value of 80 percent.

Scenario 4

The fourth scenario describes a situation that is the opposite of the first scenario. There is no wind, the temperature is high and the humidity is also high. Table 1 shows the summary of all scenario's.

TABLE 1: SUMMARY OF THE SCENARIO'S

	Temperature	Humidity	Windspeed
Scenario 1	Low	Low	High
Scenario 2	Low	Low	Low
Scenario 3	High	High	High
Scenario 4	High	High	Low

Figure 23 shows the predicted effect of water stations on the pace for the four scenarios that were presented. Scenario 1 displayed a slight decrease in pace between 0 and 3 water stations. Scenario 2 showed a slight decrease in pace between 0 and 5 water stations. Scenario 4 showed this slight decrease between 0 and 8 water stations. Scenario 3 on the other hand showed a more substantial decrease in pace when the amount of water stations increased from 6 to 9.

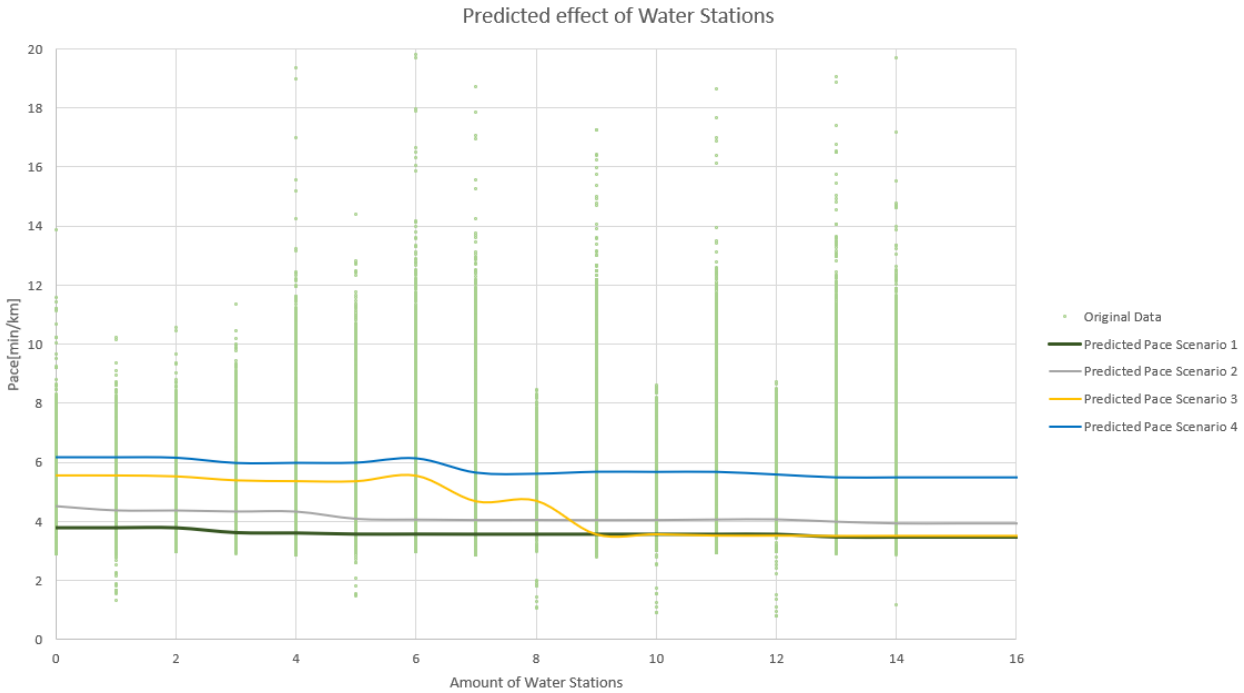

FIGURE 23: PREDICTED EFFECT OF 4 SCENARIOS

5 Conclusion

The study aimed to improve the frequency of water stations for future marathons by using a data-driven method. Several sub-questions were addressed to achieve this objective and to answer the main research question: *How can Enschede Marathon use a data-driven method to find the optimal placement frequency for water stations to achieve a sufficient performance of runners?*

Several sub-questions were answered in the introduction and literature study. Machine learning was found to be the best data driven method to model and predict the water stations of future marathons. In the results section it was concluded that in the field of machine learning, gradient boosting trees performed best in predicting the water stations. The performance of this algorithm in terms of MSE was with a score of 0.018 quite good, but the learning curve indicated some possible problems with the data.

In the machine learning predictions, a lot of features were taken into account from data that was retrieved from the Amsterdam Marathon, Rotterdam Marathon and Lisbon Marathon. The features gender, age, location, elevation, course direction, amount of water stations, temperature, humidity, wind direction and wind speed. Some of these features had a bigger impact on the pace than others. Before the machine learning was performed, a feature importance method already showed that the weather conditions Temperature, Humidity and Windspeed had the biggest impact. The results confirmed this for Temperature and Humidity with the trained machine learning algorithm. An increase in temperature would lead to an increase in pace and an increase in humidity would lead to a decrease in pace. Figure 16 also shows the combined effect of temperature and humidity, where the decrease in pace when the humidity goes up, is especially noticeable at higher temperatures. The prediction of warmer temperatures leading to runners going slower are close to the expected result based on literature. The higher humidity, especially at higher temperatures, that lead to runners going faster was not expected. Figure 22 showed the effects of wind on the pace. As predicted by the feature importance method, the algorithm also shows a big impact of the wind on the pace. There is however no clear pattern in this prediction. The effect of wind could be difficult to accurately predict as wind has both an impact on heat absorption and more challenging running conditions. Stronger winds could lead to more heat absorption for example, but could also make it more difficult for the marathon runners.

Other features had less predictive power than the weather conditions, but the results showed that some features still had an impact on the expected pace. Age was found to have a slight impact for instance. The results showed a slight increase in pace as the age goes up. This could be explained by the fact that older runners often have more experience. The distance of the marathon also had some effect on the pace. The results section showed that as the distance goes up, so the runner gets in a later stadium of the race, the pace decreases slightly. Especially the during the last 10 kilometers of the marathon, the runners are predicted to run a little faster. A possible explanation would be that runners get a motivational boost as they get further in the race.

Water stations also had an effect on the expected pace. It was shown that up to 9 water stations, the pace decreases by quite a lot. This however highly depended on temperature and that at some temperatures, the model predicted that only up to three water stations had an impact on the pace of the runner. The combined effect of water stations and location in the marathon did not show

any clear relation between the distance and the amount of water stations needed. This can be explained by the fact that the individual effect of distance on pace was also not that big. This means that water station should be placed evenly over the course of a marathon.

To look into the effects of water stations on pace further, several scenarios were created to show the results for situations with extreme weather conditions. These scenarios showed that when the temperature and humidity are high, the organization of the Enschede Marathon should place between 6 and 9 water stations evenly over the course of the marathon. The advice is to place 9 water stations for when the wind is high and 6 when the wind is lower. The advice is based on the prediction of the machine learning model that more water stations would not lead in a further decrease in pace. For cold weather conditions and less humidity, the model predicted that more than 3 to 5 water stations would not result in a decrease in pace. Again the advice would be to place 5 water stations when there is a strong wind and 3 when there is no wind. In table 2 the recommendations are summarized.

TABLE 2: RECOMMENDED AMOUNT OF WATER STATIONS FOR EACH SCENARIO

	Temperature	Humidity	Windspeed	Recommended Amount of Water Stations
Scenario 1	Low	Low	High	5
Scenario 2	Low	Low	Low	3
Scenario 3	High	High	High	9
Scenario 4	High	High	Low	6

In conclusion, this research found that temperature, humidity and wind had the most impact on the expected pace next to the amount of water stations. Distance and age also showed some impact, but this was not significant. Vihma (2010) found that from the weather conditions only temperature and humidity had a significant effect on pace, where humidity only had an effect because of the negative correlation with temperature. Lehto (2016) found a significant incline in pace after 34 years of age. This contradicts the findings of this research. Coast et al. (2004) also found that gender had an effect on pace, which was also not found in this study. These findings are summarized in table 3.

TABLE 3: SUMMARY OF EFFECTS OF FEATURES

	Effect on pace according to this study	Effect on pace according to literature
Temperature	Positive effect	Positive effect
Humidity	Some negative effect	Correlated with Temperature
Wind	Some effect	No effect
Age	Slight negative effect	Positive effect
Gender	No effect	Some effect

The main research question of the study "*How can Enschede Marathon use a data-driven method to find the optimal placement frequency for water stations to achieve a sufficient performance of runners?*" The study aimed to answer this question by using machine learning to model and

predict the impact of various factors on the pace of runners, including the number and placement of water stations. The results of the study showed that weather conditions, such as temperature, humidity, and wind, had the biggest impact on the pace of runners, but other factors, such as age and distance, also had some effect. The study recommended that water stations be placed evenly over the course of the marathon and that the number of water stations should be adjusted based on weather conditions. Overall, the study aimed to improve the frequency of water stations for future marathons by using data-driven methods to optimize their placement and number.

6. Recommendation and Discussion

The Enschede Marathon has been provided with scenarios illustrating situations where a specific number of water stations is recommended. It is advisable to adhere to these guidelines, while also considering that runners require 400-800 mL of water per hour (Cheuvront et al., 2007). Going below five water stations could make it challenging for runners to meet their hydration needs, and it is therefore suggested to avoid decreasing the number of stations even if the scenario suggests otherwise. Figure 24 displays a possible implementation of scenario 1 and 3 on the Enschede Marathon route, with the original water stations included for comparison. Remember that scenario one represent cold, non-humid and windy conditions and scenario 3 represents warm, humid and windy conditions. The icons show the water stations for these scenario's. The water stations were all placed at locations of water stations from the 2022 Enschede Marathon edition to make sure that this plan is feasible. Note that some water stations are visited twice as there is overlap in the marathon course. Note that there are other symbols added to the figure, which denote the location of music, water stations and aid posts during the 2022 edition.

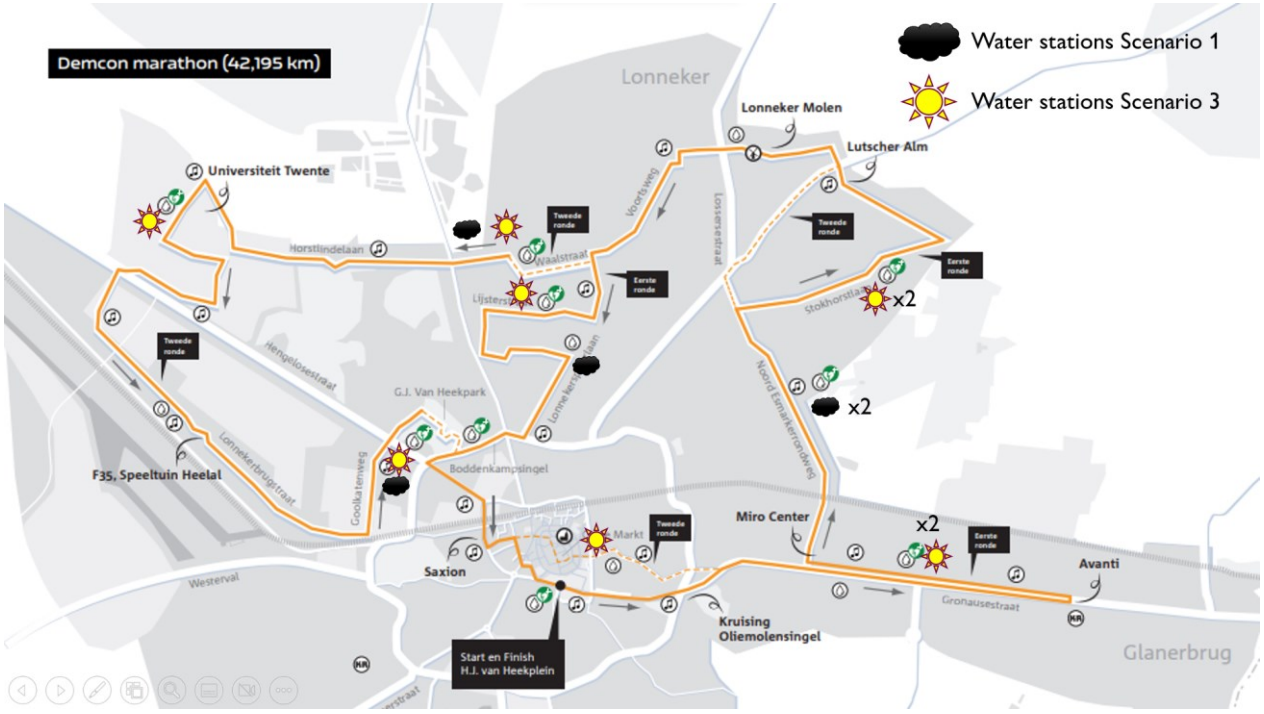

FIGURE 24: MAP OF THE ENSCHEDE MARATHON WITH TWO WEATHER SCENARIO'S

A clear limitations of this research is the underfitting of the model, which is manifested in several ways. Firstly, the low training error may suggest that the model is overfitting to the training data, but this does not guarantee its performance on new, unseen data. Secondly, the high frequency of zero errors may indicate that the model is overpredicting the majority class, which may occur due to data imbalance or the model's incapacity to distinguish the minority class's features. Lastly, the non-normal error distribution may imply underfitting, where the model is unable to capture the intricate relationships between the input features and the target variable. Thus, the model appears to underfit the data due to its low validation performance, zero errors, and non-normal error distribution. It should be emphasized that there were no studies found that use machine learning to

predict pace with similar features, hence the findings in this section cannot be fully substantiated. Nonetheless, it is highly probable that the underfitting observed in the model is due to the insufficient data variability.

To improve the model, it needs to be more intricate. Adding more features or nodes to the tree-based algorithms has not yielded significant results. In future research, a possible solution to enhance the model could be to reassess the nested cross-validation steps and test more than 50 random sets of hyperparameters per machine learning algorithm. Nonetheless, this number was set due to time limitations and may be too low to identify the optimal hyperparameters.

However, the underlying issue of underfitting most likely stems from the lack of varied data. The model may be too simplistic because the features that vary for each marathon, such as temperature, humidity, water stations, wind, and elevation, have limited variation due to the small number of marathons in the dataset. Even water stations have minimal variation across different marathon editions.

One way to resolve this, is for future research to consider adding data from more marathons. This study used data from the Amsterdam Marathon, Rotterdam Marathon, and Lisbon Marathon to model and predict the water stations for the Enschede Marathon. While these marathons are in similar regions, the findings may not be applicable to other regions or to marathons with different courses. Therefore, future studies can include data from other marathons in different regions to increase the generalizability of the findings and at the same time increase the complexity of the model.

Another way to resolve the lack of data variability and model complexity, is to add other features such as the runners' training and nutrition, as they can also affect the performance. Other interesting variables could be heart rate or sweat loss. exploring more diverse data sources and adding variables such as heart rate or sweat loss. These variables may better predict the impact of similar features on pace than the environmental features used in this study (Buchheit et al., 2010; Maxwell et al., 1996). Future studies can incorporate these factors into the models to provide more accurate predictions of water station placement.

Another recommendation for future research is to study the effects of differentiating between professional runners and recreational runners. This study did not make this distinction, but the results suggest that exploring this distinction could be valuable.

Finally, this study used a data-driven method to predict the optimal placement frequency of water stations. However, the study did not consider the runners' perceptions and preferences of water station placement. Future studies can incorporate runners' feedback and preferences into the models to provide more personalized recommendations for water station placement. By doing so, the recommendations will not only be data-driven but also more acceptable to the runners.

Contribution to Theory and Practice

This research aimed to optimize the frequency of water stations in marathons using a data-driven approach. The study utilized machine learning algorithms to predict the optimal placement of water stations for future marathons. The results showed that weather conditions, such as

temperature, humidity, and wind, had the most significant impact on the runners' pace, followed by the number of water stations. Distance and age also had some impact, but not significant.

The study provides several practical implications for marathon organizers to optimize water stations' placement. The research recommends evenly placing water stations for up to 9 stations for high temperature and humidity, while for low temperature and humidity, it recommends 3 to 5 water stations. The placement of water stations should also depend on wind speed, and organizers should place 6 water stations when the wind speed is high and 9 water stations when the wind speed is low.

The research also contributes to the theoretical understanding of the factors that affect runners' pace in marathons. The study found that weather conditions were the most critical predictors of pace, which was consistent with previous research. However, the research found that humidity had a positive effect on pace, which was not expected. The study also contradicts previous research that found gender and age had a significant impact on pace.

In conclusion, this study provides practical recommendations for optimizing water stations' placement in future marathons using a data-driven approach. It also contributes to the theoretical understanding of the factors that affect runners' pace in marathons.

References

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, *3*(1), 1-122.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Buchheit, M., Chivot, A., Parouty, J., Mercier, D., Al Haddad, H., Laursen, P., & Ahmaidi, S. (2010). Monitoring endurance running performance using cardiac parasympathetic function. *European journal of applied physiology*, *108*, 1153-1167.

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1). Andriy Burkov Quebec City, QC, Canada.

Casa, D. J., Clarkson, P. M., & Roberts, W. O. (2005). American College of Sports Medicine roundtable on hydration and physical activity: consensus statements. *Current sports medicine reports*, *4*(3), 115-127.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.

Cheuvront, S. N., & Haymes, E. M. (2001). Ad libitum fluid intakes and thermoregulatory responses of female distance runners in three environments. *Journal of sports sciences*, *19*(11), 845-854.

Cheuvront, S. N., Montain, S. J., & Sawka, M. N. (2007). Fluid replacement and performance during the marathon. *Sports medicine*, *37*(4), 353-357.

Coast, J. R., Blevins, J. S., & Wilson, B. A. (2004). Do gender differences in running performance disappear with distance? *Canadian Journal of Applied Physiology*, *29*(2), 139-145.

Dancaster, C., & Whereat, S. (1971). Fluid and electrolyte balance during the comrades marathon. *South African Medical Journal*, *45*(2), 147-150.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378.

Goldberg, D. E., Korb, B., & Deb, K. (1989). Messy genetic algorithms: Motivation, analysis, and first results. *Complex systems*, *3*(5), 493-530.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41). Springer.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Haykin, S. S. (2009). Neural networks and learning machines/Simon Haykin. In: New York: Prentice Hall.

Historie - Enschede Marathon. (2022). https://www.enschedemarathon.nl/historie/

Knechtle, B., Di Gangi, S., Rüst, C. A., Villiger, E., Rosemann, T., & Nikolaidis, P. T. (2019). The role of weather conditions on running performance in the Boston Marathon from 1972 to 2018. *PloS one*, *14*(3), e0212797.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, *6*(1), 1-15.

Lara, B., Salinero, J. J., & Del Coso, J. (2014). The relationship between age and running time in elite marathoners is U-shaped. *Age*, *36*(2), 1003-1008.

Lehto, N. (2016). Effects of age on marathon finishing time among male amateur runners in Stockholm Marathon 1979–2014. *Journal of Sport and Health Science*, *5*(3), 349-354.

Margaria, R., Cerretelli, P., Aghemo, P., & Sassi, G. (1963). Energy cost of running. *Journal of applied physiology*, *18*(2), 367-370.
Maughan, R. (2003). Impact of mild dehydration on wellness and on exercise performance. *European journal of clinical nutrition*, *57*(2), S19-S23.
Maxwell, N., Aitchison, T., & Nimmo, M. (1996). The effect of climatic heat stress on intermittent supramaximal running performance in humans. *Experimental Physiology: Translation and Integration*, *81*(5), 833-845.
Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.
Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62.
Nested Cross Validation. (2020). https://vitalflux.com/python-nested-cross-validation-algorithm-selection/
Noakes, T. (2003). Fluid replacement during marathon running. *Clinical Journal of Sport Medicine*, *13*(5), 309-318.
Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74-78.
Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.
Reitermanova, Z. (2010). Data splitting. WDS,
Sedeaud, A., Marc, A., Marck, A., Dor, F., Schipman, J., Dorsey, M., Haida, A., Berthelot, G., & Toussaint, J.-F. (2014). BMI, a performance parameter for speed improvement. *PloS one*, *9*(2), e90183.
Shapiro, Y., Pandolf, K. B., & Goldman, R. F. (1982). Predicting sweat loss response to exercise, environment and clothing. *European journal of applied physiology and occupational physiology*, *48*(1), 83-96.
Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.
Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.
Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 1-8.
Vihma, T. (2010). Effects of weather on the performance of marathon runners. *International journal of biometeorology*, *54*(3), 297-306.
Wang, S.-C. (2003). Artificial neural network. In *Interdisciplinary computing in java programming* (pp. 81-100). Springer.
Williams, J., Brown, V. T., Malliaras, P., Perry, M., & Kipps, C. (2012). Hydration strategies of runners in the London Marathon. *Clinical Journal of Sport Medicine*, *22*(2), 152-156.
Wyndham, C., & Strydom, N. (1969). The danger of an inadequate water intake during marathon running. *South African Medical Journal*, *43*(29), 893-896.
Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.

Appendix

Appendix A: Hyperparameter ranges

Algorithm	Parameter	Range
Random Forest	'n_estimators'	[50, 100, 150, 200, 250, 300, 400, 500]
Random Forest	'max_features'	['auto', 'sqrt']
Random Forest	'max_depth'	[None, 5, 8, 10, 12, 16, 20]
Random Forest	'min_samples_split'	[2, 4, 6, 8, 10]
Random Forest	'min_samples_leaf'	[1, 3, 5, 7]
Linear Regression	'alpha'	Range[0, 1, 0.01]
Neural Network	learning_rate'	['constant', 'invscaling', 'adaptive']
Neural Network	batch_size'	[10,25,40,55,70]
Neural Network	max_iter'	[5,50,100,150,250,350,500,650]
Neural Network	validation_fraction'	[0.01,0.1,0.2]
Gradient Boosting Trees	'n_estimators'	[50, 100, 150, 200, 250, 300, 400, 500]
Gradient Boosting Trees	'max_features'	['auto', 'sqrt']
Gradient Boosting Trees	'max_depth'	[None, 5, 8, 10, 12, 16, 20]
Gradient Boosting Trees	'min_samples_split'	[2, 4, 6, 8, 10]
Gradient Boosting Trees	'min_samples_leaf'	[1, 3, 5, 7]
Decision Tree	'max_features'	['auto', 'sqrt']
Decision Tree	'max_depth'	[None, 5, 8, 10, 12, 16, 20]
Decision Tree	'min_samples_split'	[2, 4, 6, 8, 10]
Decision Tree	'min_samples_leaf'	[1, 3, 5, 7]

FIGURE 25: HYPER PARAMETER RANGES

Appendix B: Learning curves

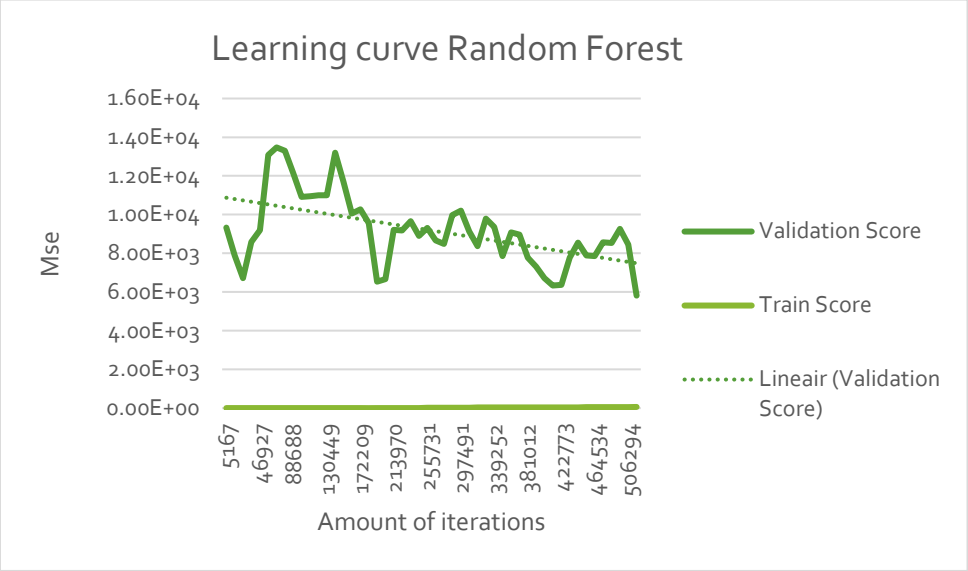

FIGURE 26: LEARNING CURVE RANDOM FOREST

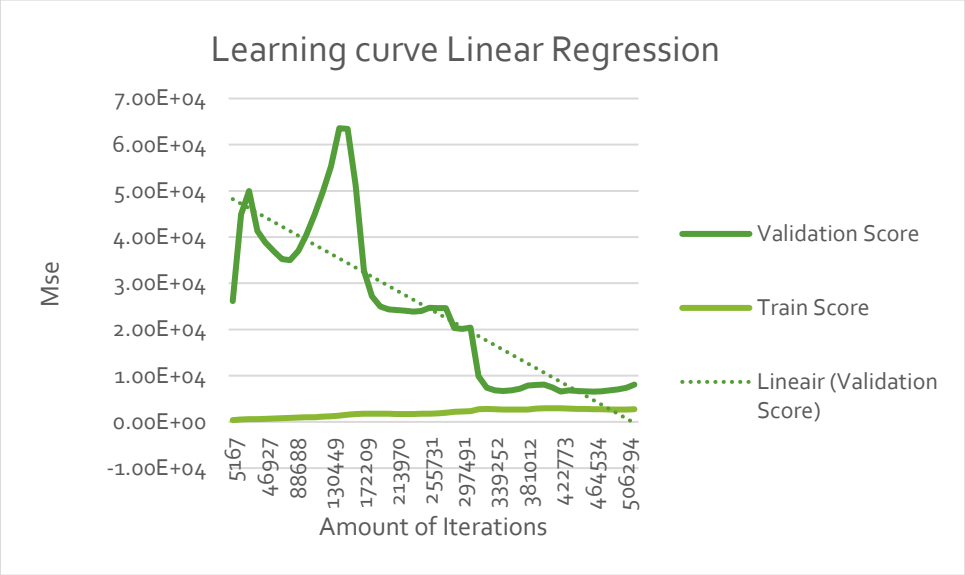

FIGURE 27: LEARNING CURVE LINEAR REGRESSION

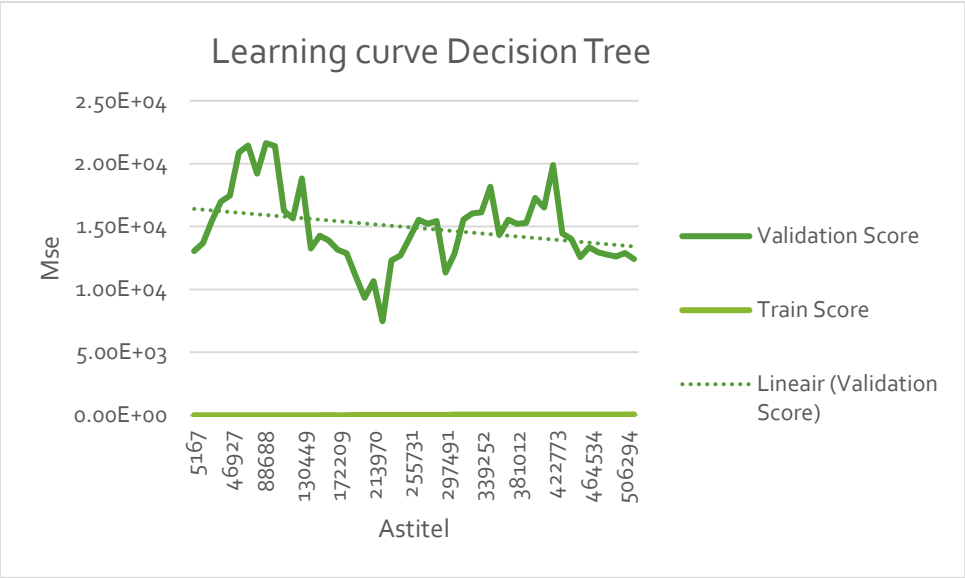

FIGURE 28: LEARNING CURVE DECISION TREE

Appendix C: Python Code

```python
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import pandas as pd
import pickle

PATH = "C:\Program Files (x86)\chromedriver.exe"
driver = webdriver.Chrome(PATH)

url ="https://results.sporthive.com/events/685374155229023872O/races/479865"

with open("test1", "rb") as fp:   # Unpickling
   Runner_list = pickle.load(fp)

for Runner in Runner_list:
    Bibrunner = Runner["Bib"]
    Newurl= url + "/bib/" + Bibrunner
    driver.get(Newurl)
    counter=0
    try:
        Wait = WebDriverWait(driver, 10).until(
            EC.presence_of_element_located((By.TAG_NAME, 'tr.ng-scope')))
        Chiptimesrunner = driver.find_elements_by_xpath('//div[2]/table/tbody/tr')
        for Chiptimerunner in Chiptimesrunner:
            counter += 1
            Currentchiptime = Chiptimerunner.find_element_by_tag_name("span.ng-binding.ng-scope")
            Location = Chiptimerunner.find_element_by_tag_name("td.col-is-name.ng-binding").text

            Pace = "Pace" + str(counter)
            Locationname = "Location" + str(counter)
            Runner[Pace] = Currentchiptime
            Runner[Locationname] = Location
    except:
        driver.quit()

df = pd.DataFrame(Runner_list)

Excelfile = str(url)+".xlsx"
writer = pd.ExcelWriter("Amsterdam2021.xlsx", engine='xlsxwriter')
df.to_excel(writer, sheet_name='Test1', index=False)
writer.save()

driver.quit()
```

FIGURE 29: WEB SCRAPER

```
import pickle
from sklearn.model_selection import KFold
from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import GradientBoostingRegressor

with open("C:/Users/alber/OneDrive/Documenten/IEM/Master Thesis/Final data/Alldata4.0", "rb")
   Alldata = pickle.load(fp)
#create real datasets
dfX = Alldata.drop('Pace (s/km)',axis=1)

dfy = Alldata['Pace (s/km)']
X = dfX.to_numpy()
Y = dfy.to_numpy()


# configure the cross-validation procedure
cv_outer = KFold(n_splits=3, shuffle=True, random_state=0)
# enumerate splits
outer_results = list()
outer_bestscore = list()
outer_bestparameters = list()
for train_ix, test_ix in cv_outer.split(X):
    # split data
    X_train, X_test = X[train_ix, :], X[test_ix, :]
    y_train, y_test = Y[train_ix], Y[test_ix]
   # configure the cross-validation procedure
   cv_inner = KFold(n_splits=3, shuffle=True, random_state=0)
   # define the model
   model = GradientBoostingRegressor(random_state=0) #!!!
   # define search space
   space = dict() #!!!
   space['n_estimators'] = [50,100,150,200,250,300,400,500]
   space['max_features'] = ['auto', 'sqrt']
   space['max_depth'] = [None,5,8,10,12,16,20]
   space['min_samples_split'] = [2,4,6,8,10]
   space['min_samples_leaf'] = [1,3,5,7]

   # define search
   search = RandomizedSearchCV(model, space, n_jobs=-1,
                               scoring='neg_mean_squared_error', cv=cv_inner,
                               refit=True, n_iter = 50)
   # execute search
   result = search.fit(X_train, y_train)
   # get the best performing model fit on the whole training set
   best_model = result.best_estimator_
   # evaluate model on the hold out dataset
   yhat = best_model.predict(X_test)
   # evaluate the model
   acc = (y_test - yhat)**2
   # store the result
   outer_results.append(acc)
   # report progress
   bestscore = result.best_score_
   bestparameters = result.best_params_
   outer_bestscore.append(bestscore)
   outer_bestparameters.append(bestparameters)
   print("Round finished")
```

FIGURE 30: NESTED CROSS VALIDATION GRADIENT BOOSTING TREES