Smart Hydration Planning Based on Workout and Weather Data

MSc Research Project
Data Analytics

Rebecca Pina
Student ID: x23436786

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rebecca Pina
Student ID:	x23436786
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Arjun Chikkankod
Submission Due Date:	11/08/2025
Project Title:	Smart Hydration Planning Based on Workout and Weather Data
Word Count:	6032
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Smart Hydration Planning Based on Workout and Weather Data

Rebecca Pina
x23436786

Abstract

Hydration is essential for the proper functioning of the human body and essential for recreational athletes during physical exercise. However, many of these athletes lack access to personalized hydration strategies before and during workouts. While some smartwatches offer water intake reminders, these are often generic and not tied to workouts or individual physiological history. This project aims to fill the hydration gap for recreational athletes by developing a hydration assistant that provides personalized recommendations based on planned workouts, the user's physical attributes (e.g., age, weight, gender), exercise duration, intensity, and weather data such as temperature. Unlike current solutions, which rely on smartwatches with expensive monthly subscriptions, this assistant uses machine models trained on a public and private exercise dataset, combined with sweat loss formulas and environmental information to estimate the athlete's water needs. The system then provides recommendations through a web interface. This solution is suitable for beginner athletes who want to improve performance and avoid dehydration without the need for high-end smartwatches.

Keywords: Hydration, Machine Learning, Fitness Technology, Exercise Planning.

1 Introduction

Hydration is essential for performance, recovery, and, most importantly, injury prevention in athletes. Particularly for athletes who engage in endurance activities such as long-distance running or cycling, maintaining hydration is essential not only for performance but also for safety. Dehydration has been shown to increase body temperature, impair thermoregulation, and increase the risk of heat-related illness, especially in hot and humid environments. Casa et al. (2019).

Despite the importance of adequate hydration, studies such as Judge et al. (2021) have shown that athletes lack sufficient knowledge about how to properly hydrate. Furthermore, thirst alone cannot be used to assess an athlete's hydration level, as significant fluid loss can occur before thirst is even noticed.

The need for hydration planning is especially important for athletes who train in locations without access to hydration stations or who exercise for long periods. This is the case for runners and cyclists participating in activities longer than 10km, who are at high risk of fluid and electrolyte loss, requiring not only water intake planning but also salt replacement. These nuances demonstrate the importance of personalized guidance guided by the athlete's context, including the type of sport practiced.

Advances in wearable technology can address this gap, with a large number of devices capable of monitoring hydration levels, including indicators such as skin conductivity, sweat, and thermal properties. Belabbaci et al. (2025) research categorized existing sensors by type, including conduction, capacitance, optical, and microwave-based, each with different strengths and limitations.

Similarly, Benavides et al. (2021) developed a useful system that monitors hydration and UV exposure through a mobile app. While these technologies may be promising, they remain inaccessible to most of the population due to their high cost and the time required between research and commercialization, in addition to the need for constant use of the wearable and dependence on accuracy and calibration.

In response to these challenges, this project proposes the development of a non-invasive hydration assistant powered by supervised machine learning. Unlike existing solutions that rely on wearable sensors and constant physical data, the proposed system offers personalized water intake recommendations based on scheduled physical activity and ambient weather conditions.

The model is trained using a combination of public data from anonymized and private gym users, incorporating specific attributes (such as age, gender, weight, height, type of exercise performed, duration of exercise, and amount of water consumed) along with contextual data such as temperature.

The main objective of this project is to prove accessible and personalized hydration planning for beginner/amateur athletes, eliminating the need for expensive sensors and apps with subscriptions or live connections.

The report is structured in six sections, the second of which is a review of the existing literature on hydration science, wearable monitoring technologies, and machine learning applications in fitness analysis. Section 3 outlines the research methodology, including data processing, feature engineering, and model evaluation. Section 4 describes the system design and implementation. Section 5 presents the results and performance analysis. Finally, Section 6 summarizes the main findings and outlines areas for future research.

2 Related Work

2.1 Hydration and Its Physiological Importance

The Casa et al. (2019) study emphasizes on the need for fluids for training, competition, and athlete recovery, and the impacts that seasonal changes in the environment can pose, challenging even well-trained athletes. It highlights the risks of dehydration and its consequences for athlete performance. The 2011 International Olympic Committee reached a consensus regarding hydration: athletes should be hydrated before exercise, and if the activity lasts more than two hours, include sodium, especially if sweat loss is high.

Expanding on this, Chodkowski (2024) explores hydration's role in maintaining physiological balance, highlighting its role in fluid balance and physiological functions such as thermoregulation, nutrient transport, and muscle function. It is important to note that dehydration, even moderate, can cause cramps and problems with thermal regulation, leaving athletes susceptible to heat-related illnesses, including heat exhaustion and heat stroke. But it doesn't stop there; dehydration also affects cognition, increasing mental fatigue and declining concentration, focus, and reaction time. This highlights the need for personalized hydration tailored to the individual needs of each athlete.

According to Papaoikonomou et al. (2025), who focused on hydration among adolescents and children, he found that among athletes, 81% of studies demonstrated dehydration, while 69% of studies observed dehydration in non-athlete children. This demonstrates the need for education on proper hydration for children and adolescents, especially those who engage in physical exercise.

Holland-Winkler and Hamil (2024), on the other hand, researched hydration in professionals such as firefighters, with reports showing that approximately 90% of firefighters are continually dehydrated. He noted that cardiac death accounted for the majority of firefighter deaths between 1997 and 2019, a finding that may be related to dehydration. Dehydration increases stress on the heart, leading to an increased risk of developing cardiovascular disease. The Institute of Medicine recommends a daily intake of 3.7 liters for men and 2.7 liters for women and not exceeding 8.5 liters per day.

2.2 Wearable Hydration Monitoring Technologies

Gray et al. (2023) examines emerging hydration monitoring technologies, ranging from bio-impedance and electromagnetic probes to infrared and Raman spectroscopy, as well as wearable sensors that assess sweat or saliva composition. While many of these approaches demonstrate theoretical sensitivity to water content and are viable for routine measurements, the authors point out that few have undergone rigorous validation to ensure diagnostic accuracy. Bringing these emerging hydration monitoring technologies to the sports realm, not all options are viable for use. Gray et al. (2023) study also highlights the need for robust clinical evaluation to translate these technologies into effective real-world tools.

While Belabbaci et al. (2025) developed a lengthy study categorizing sensor technologies into electrical, optical, thermal, microwave, and multimodal, including eight commercial products in his analysis. He detailed their strengths and limitations, such as the dominance of electrical sensors due to their easy integration with wearables. BIA (Bioelectrical Impedance Analysis) offers good metrics but is quite sensitive to variables such as skin type. Optical sensors, for example, demonstrate high accuracy, while multimodal sensors are an emerging trend due to their accuracy. Sweat analysis using electrochemical sensors shows promise, but we encounter challenges regarding stability and sweat variability.

Sreeharsha et al. (2024), on the other hand, focused on the use of machine learning, its capabilities, limitations, and future work. Machine learning techniques such as Random Forest, Decision Trees, and Deep Neural Networks increase the accuracy of hydration prediction using data from ECG, GSR, sweat sensors, etc. However, the lack of public databases, sensor limitations, and limited testing in diverse population groups have proven to be a challenge. While machine learning and multi-sensor systems show potential for personalized hydration monitoring, practical application remains difficult due to data scarcity, system complexity, and the lack of benchmarks.

Włoch et al. (2025) presents an overview of modern technologies used in sports training, including sensors/wearable devices, GPS trackers, heart rate monitors, motion sensors, and AI-powered feedback systems. These technologies have been proven to increase training efficiency by enabling real-time performance monitoring, enabling injury prevention, and providing personalized feedback.

The study highlights the value of data-driven systems for adjusting training loads, optimizing biomechanics, and improving athlete recovery—principles that can be extended to hydration planning. It is important to highlight the potential limitations, including over-reliance on technology and ethical concerns. This work supports the integration of personalized, accessible, and non-invasive technologies in sports settings, emphasizing the need for hydration systems that provide performance-enhancing feedback without costly and continuous physical monitoring.

2.3 Artificial Intelligence in Hydration Prediction

In recent literature on machine learning in healthcare, specifically in the health sector, the study by Ramírez. et al. (2024) stands out. He conducted a literature review focused on the evaluation, monitoring, and maintenance of machine learning models in practical healthcare settings. His findings revealed that although machine learning applications are increasingly used for diagnosis, prognosis, and treatment, there are gaps, such as the lack of system model evaluation and post-implementation protocol maintenance. Models in healthcare must be adaptable to new input distributions and operational contexts. In hydration monitoring, this is demonstrated by the need for mechanisms to detect changes in the user's physical fitness, climate variations, or device sensor behavior.

Other systematic reviews, such as that by Gupta et al. (2025), which analyzed quantum machine learning (QML) applications for digital health, found that most studies relied on ideal simulations and failed to test models under realistic conditions, such as in environments with variable data quality. While this review focused on QML, its conclusions also apply to AI-based healthcare systems, demonstrating the need for evaluations that test models outside of ideal conditions. This emphasis on real-world performance, scalability, and reproducibility is relevant to the development of AI-based hydration assistants. These systems must be prepared for diverse input conditions, environmental factors, and potential sensor inaccuracies.

The need for privacy-preserving machine learning is particularly relevant in healthcare applications, due to the availability of data and regulatory requirements that often limit the predictability of centralized training. Teo et al. (2024) conducted a systematic review of federated learning (FL) in healthcare, analyzing over 600 studies predominantly from radiology and clinical medicine. They found that FL enables collaborative training of models across different institutions without the need for raw data transfer, thus addressing a fundamental limitation of traditional centralized machine learning.

These insights are important for AI-based hydration monitoring systems, which can benefit from training from multiple sources, such as wearable devices, different companies, or sports clubs, while respecting user privacy and data protection regulations such as GDPR. One of the current limitations for training hydration assistants is the lack of broader and more diverse data without compromising personal health information. Furthermore, the governance, infrastructure, and explainability considerations identified in the Teo et al. (2024) study are structured around the long-term implementation challenges for hydration systems, reinforcing the need for safety, scalability, and ethical management in sports and health technologies.

Prova (2024) explores the application of advanced machine learning regression techniques for predictive analysis in health insurance, comparing decision tree, random forest, XBG, and KNN regressors on a dataset with over 70,000 records. The decision tree regressor (R^2 = 0.85) stood out as the best-performing model due to its interpretability and ability to effectively capture variation, followed by the random forest regressor (R^2 = 0.83). XBG and KNN achieved lower predictive accuracy.

In addition to model evaluations, the research highlights important preprocessing steps, such as outlier removal using the interquartile range method, label encoding for categorical variables, and exploratory data visualization to identify distributions and feature patterns. Although the study's context is health insurance cost prediction, its methodological approach offers relevance to the AI-based hydration system. An architected model comparison provides a framework for selecting algorithms capable of delivering high predictive accuracy and interpretability. Furthermore, the preprocessing step described by Prova (2024) can be adapted to the hydration domain to ensure a clean and representative dataset for model training.

2.4 Machine Learning Application in Sports

Focused on sports medicine, Pareek et al. (2024) provides an overview of how machine learning and artificial intelligence are advancing sports medicine, emphasizing applications in injury prediction, rehabilitation, performance optimization, and image interpretation. It demonstrates that machine learning models can help anticipate musculoskeletal injuries based on load, movement, and recovery metrics, emphasizing the growing role of AI in creating personalized, data-driven interventions that enhance athletic performance and reduce injury risk. Similar models predict dehydration risk based on planned exercise, personal physiological data, and environmental conditions.

Galekwa et al. (2024) explored the role of machine learning in sports betting, identifying key techniques such as support vector machines, random forests, and neural networks used in various sports such as basketball, soccer, tennis, and cricket. His findings highlighted how machine learning models improve predictive accuracy and enable risk management for both bookmakers and bettors. Although the focus of Galekwa et al. (2024) study is different from this project, the Hydration System based on supervised machine learning can be used as a benchmark for evaluating the performance and applicability of machine learning models in prediction-related tasks.

Richter et al. (2024) presents a critical overview of the opportunities and challenges associated with implementing machine learning in Sports Science. They emphasize that, although ML models have shown promise in several areas, the success of these applications depends heavily on methodological rigor. The authors detailed the main phases of the ML pipeline—data acquisition, feature engineering, and evaluation model selection—highlighting issues such as noise in inter-laboratory data, overfitting, and the misuse of accuracy as a performance metric. This provides a valuable perspective for framing the practical constraints and ethical considerations discussed in this thesis.

Leckey et al. (2025) converted a study where machine learning is applied to predicting sports injury risk, identifying tree-based algorithms—Random Forest and XGBoost—as the most frequently applied and best-performing approaches. He revealed that in 60% of cases, solutions using these models outperformed other methods. He highlighted the relevance of evaluations, such as the area under the receiver operating curve (AUC), reported in 71% of studies, providing better insight into model performance in imbalanced datasets compared to accuracy alone.

These metrics are directly relevant to AI-based hydration prediction, where datasets can be imbalanced and where models that do not impose significant machine processing complexity are more viable options for real-world applications. Leckey et al. (2025) also cautions against models that achieve high statistical performance but lack actionable specificity, such as poorly defined outcome variables.

2.5 Ethical and Practical Considerations

Kotowicz et al. (2025) presents a comprehensive overview of the applications of artificial intelligence in various areas of medicine, emphasizing the principles of transparency, equity, and "augmented intelligence"—where AI complements, rather than replaces, human decision-making. A central topic is the use of multimodal approaches, in which data from multiple sources are combined to improve prediction accuracy and personalization, enabling earlier and more accurate diagnoses in ophthalmology and radiology, for example. The connection with AI-based hydration systems, bringing a multimodal view to data from different sources, is applicable, potentially producing more accurate and individualized fluid intake recommendations.

While Ramírez. et al. (2024) and Kotowicz et al. (2025) explore the role of artificial intelligence in medicine, their focuses and intentions are different, but they provide complementary insights for this thesis. Ramírez. et al. (2024) focuses on the lifecycle of machine learning in healthcare, its gaps, and post-implementation needs, thus ensuring optimal long-term performance. In contrast, Kotowicz et al. (2025) offers a broader and more specific survey of applications in key areas of medicine. It demonstrates the importance not only of building large AI systems but also of ensuring their scalability, explainability, and safe integration into clinical settings—principles directly applicable to design.

Padulo (2025) presents a perspective on sports and health sciences, highlighting the integration of technology, psychology, and data analytics to personalize performance monitoring and rehabilitation. His report highlighted the growing role of wearables, AI, and big data in precision sports care, particularly in tailoring interventions that address physical and wellness aspects. This interdisciplinary perspective is deeply aligned with the development of the Smart Hydration System, as it validates the use of data-driven approaches to improve health outcomes. Padulo (2025) highlights the potential risks associated with the diffusion of wearables and AI, including data privacy, unequal access, and over-reliance on automated systems for decision-making. This aligns with the Teo et al. (2024) vision, which emphasizes privacy-preserving strategies for training AI models in institutions. Similarly, Ramírez. et al. (2024) addresses the post-implementation phase, noting the lack of governance mechanisms that ensure reliability over time.

3 Methodology

This study adopts an applied quantitative research design, using a predictive machine learning approach to address a topic in the field of exercise and health. The goal is to build a practical, data-driven system that enables better hydration planning for beginning athletes, particularly those without access to advanced wearable technologies. Rather than traditional hydration tools that offer only generic recommendations/alerts or rely on continuous physiological monitoring, this research focuses on predicting the amount of water an athlete needs to consume during a planned training session.

This approach is based on user-provided inputs, making it especially useful for amateur or recreational athletes who do not have access to advanced monitoring equipment. The approach follows the framework of the Cross-Industry Standard Process for Data Mining (CRISP-DM). The project's methodology choice is based on the sports science literature, which highlights the strong correlation between hydration needs and measurable factors such as exercise type, intensity, ambient temperature, and body mass Casa et al. (2019).

Machine learning has been increasingly adopted in sports performance analysis Pareek et al. (2024), making it a suitable methodological basis for this study. The model is designed to provide users with personalized hydration estimates based on the characteristics of the planned exercise, the user's physiological characteristics, and contextual environmental data.

By combining a public dataset, personal exercise logs collected with a Polar watch, and weather data (temperature), the system applies supervised machine learning techniques to estimate required fluid intake. This design allows the system to generalize to different types of activities and weather conditions, while remaining accessible to users without wearable sensors.

3.1 CRISP-DM Process Steps

3.1.1 Data Sources

Two main datasets were used to train the hydration prediction model:

- Gym Members Exercise Dataset (Kaggle):

A publicly available dataset containing structured information on gym workouts, including age, gender, weight (kg), max BPM, avg BPM, resting BPM, session duration, calories burned, workout type, fat percentage, water intake (liters), workout frequency, experience level, BMI.

- Personal Workout Logs:

A dataset created by the author using her own real workout history. This includes personal details such as age, weight, height, and specifics of each workout session (duration, type, pace, perceived intensity). The data used does not include real-time biometric data or third-party health data.

3.1.2 Data Preprocessing

The personal training log dataset was in JSON format and split into multiple files for a single workout. First, all 292 JSON files were merged into a single JSON file, which was then converted to a CSV file. The Academy Member Training Dataset (Kaggle) was aggregated, forming a single CSV file with the data ready for preprocessing. The preprocessing phase comprises the following steps:

- Data Cleaning:

Removal of incomplete or inconsistent records, such as columns with missing values or columns unnecessary for processing (such as: Workout_Frequency (days/week)). Standardization of units (minutes → seconds, height in cm, weight in kg).

- Data Consistency Adjustments:

The Gender and Exercise_Type columns have been normalized to uppercase to avoid category duplication caused by inconsistent capitalization. Variants of the same exercise have been mapped to a standard label (Strength -¿ Strength_Training).

- Outliers:

Some extreme values in performance-related characteristics (duration, distance, pace) are expected for high-level training, due to the nature of the data.

- Data Export:

After cleaning and normalization, the processed dataset was stored in combined_data1.csv. This file served as input for the subsequent feature engineering phase and the model training process.

3.1.3 Feature Engineering

From the cleaned dataset, a function using the sweat loss formula was applied to calculate the amount of water needed during exercise. This value was stored in a column, as one of the original datasets did not have this measurement.

The following additional features were designed to improve predictive performance: BMI – calculated from height and weight, Exertion Score – derived from exercise type, duration, and pace, Long Distance Flag – boolean feature indicating long-duration endurance sessions, Ambient Temperature – randomly generated within realistic ranges during training to improve generalization; provided by the user during implementation.

These designed variables were added to the dataset to create the final feature matrix (X) and target vector (y).

3.1.4 Model Development and Training

The model development process followed the CRISP-DM modeling phase, where multiple supervised regression algorithms were implemented and evaluated to identify the best-performing approach for hydration volume prediction. The final predictive model was implemented as a Stacking Regressor, combining: Random Forest Regressor, XGBoost Regressor, Ridge Regression.

A ColumnTransformer was used in the pipeline to scale numerical features and one-hot encode categorical variables (Gender, Exercise Type, Location). The model was trained on 80% of the dataset, with 20% reserved for testing. Cross-validation was applied to ensure generalizability and to tune hyperparameters.

- Baseline Model:

Linear Regression was used as baseline due to its simplicity and interpretability. It assumes a linear relationship between features and target, providing a good benchmark performance.

- Tree-Based Model:

Random Forest was selected for its ability to handle non-linear relationships and capture complex feature interactions without the need for extensive preprocessing. XGBoost Regressor was included for its proven effectiveness on tabular datasets and its gradient boosting framework.

- Regularized Linear Models:

Lasso regression was applied for its benefits in feature selection, as it can reduce irrelevant coefficients to zero, potentially simplifying the model. Ridge regression was used to address multicollinearity and stabilize coefficient estimates, reducing the risk of overfitting.

- Instance-Based Model:

K-Nearest Neighbors (KNN) was tested to explore a nonparametric instance-based learning approach that predicts based on the closest training examples in the feature space.

- Ensemble Learning:

After individual model evaluations, a joint stacking strategy was adopted to combine the strengths of the different models. The final Stacking Regressor included: Random Forest Regressor, XGBoost Regressor and Ridge Regression.

- Model Selection Criteria:

The models were compared using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. The Stacking Regressor consistently outperformed the individual models across all metrics, providing lower error rates and better explanation of variance, making it the final model for deployment.

4 Design Specification

The proposed solution follows a machine learning architecture that separates the layers of data ingestion, preprocessing, feature engineering, model inference, user interaction and Streamlit web application. This structure facilitates maintenance and scalability without the need for significant reengineering.

4.1 Core Framework

4.1.1 Architecture:

CRISP-DM methodology adapted for hydration prediction, ensuring a structured approach from data acquisition to implementation.

4.1.2 Frameworks and Libraries:

- Pandas/NumPy for data processing and preprocessing.
- Scikit-learn for training, evaluation, and model pipeline management.
- XGBoost for a high-performance gradient boosting model, cited in the literature as one of the options for prediction.
- Streamlit to deploy a lightweight, interactive, easy-to-use web application, making the system lightweight and accessible to all types of smartphones.

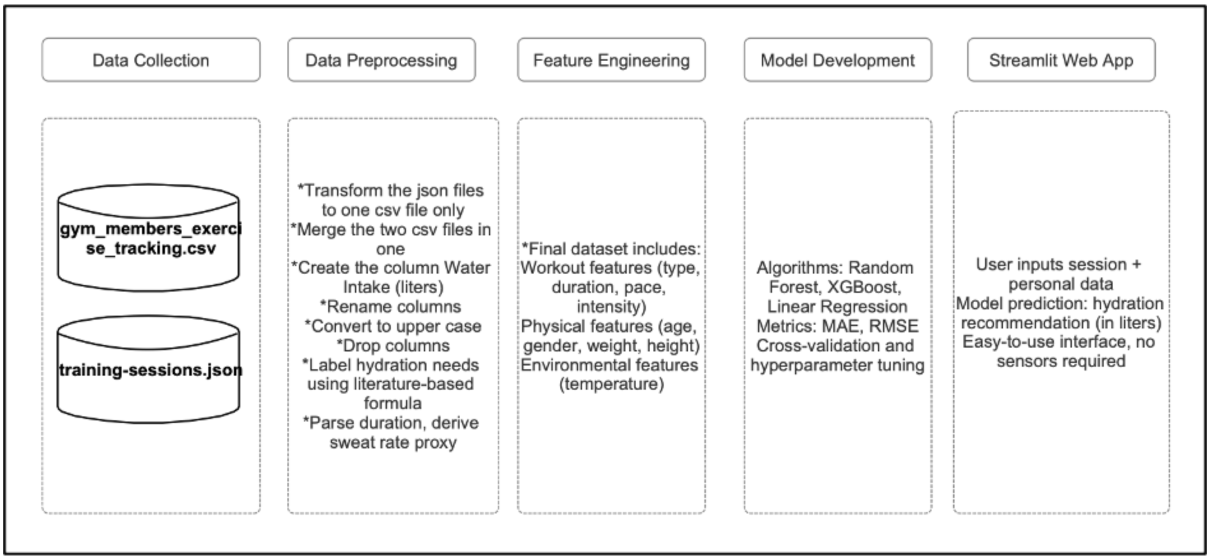

Figure 1: Project Design

4.2 System Requirements

4.2.1 Input:

Planned exercise details (type, duration, distance—for locomotion exercises such as walking, running, and treadmill running), physiological data (age, sex, height, weight), and ambient temperature.

4.2.2 Output:

Predicted water intake in liters for the planned training session.

4.2.3 Accessibility:

The system operates independently of wearable devices or sensor integration to ensure broad usability.

4.2.4 Performance Target:

Achieve RMSE less than or equal to the benchmark established during model evaluation (base model).

4.3 Architecture

This architecture was deliberately designed to be flexible and accessible, allowing the model to function effectively in environments without wearable device integration and without the need for continuous personal data. The use of a stacked ensemble ensures robust prediction accuracy, leveraging the strengths of multiple learning algorithms, while the CRISP-DM framework provides a systematic development and evaluation framework. The main focus of the system and this architecture is to maintain a simple yet efficient approach that is accessible to everyone, whether they are new to sports or experienced.

4.3.1 Data Ingestion Layer

For the training phase: Reads and validates input from CSV datasets. For the deployment phase: Accepts form input from the user interface.

4.3.2 Preprocessing Layer

Handles missing or inconsistent values, normalizes numerical variables and encodes categorical features to match the training schema.

4.3.3 Feature Engineering Layer

Here is the sweat loss formula, calculates BMI, exertion score and other derived metrics to increase predictive power.

4.3.4 Prediction Layer

Utilizes a Stacking Regressor combining Linear Regression, Random Forest, and XGBoost models.

4.3.5 Presentation Layer

Built in Streamlit, enabling real-time interaction. Displays predicted hydration needs.

5 Implementation

The final implementation provides a fully operational Intelligent Hydration Planning System capable of generating personalized water intake recommendations based on planned workout parameters, physiological attributes, and weather conditions. The system produces the following outputs:

- Transformed Data: Input values are cleaned, standardized, and encoded to match the schema of the trained model.
- Machine Learning Predictions: Water intake recommendations in liters, tailored to the user's planned exercise session.
- Interactive User Interface Output: Hydration recommendations displayed.

To enable supervised learning, a suggested water-intake target was calculated for each workout using a formula based on the principles of sweat loss and aligned with common sports science guidelines (approximately 0.4–0.8 L/h, adjusted for workload and environmental heat).

Hydration Target Construction

To create the supervised label, a sweat-loss–informed water-intake target was computed from calories and adjusted by physiological and environmental factors. The formulas presented here were based on scientific studies on hydration, always paying attention to physiological limits and human health.

Let d_h be the session duration in hours, w the body mass (kg), T the ambient temperature (°C), and C the calories burned.

Intensity factor.

$$m_I \in \{1.0\ (\text{low}),\ 1.2\ (\text{moderate}),\ 1.5\ (\text{high})\}.$$

Base sweat-loss from calories.

$$\text{base\_loss} \;=\; \frac{C}{700}.$$

Temperature factor (corrected bands).

$$\text{temp\_fac} \;=\; \begin{cases} 1.00, & T < 15, \\ 1.10, & 15 \le T \le 25, \\ 1.25, & 25 < T \le 32, \\ 1.40, & T > 32. \end{cases}$$

Gender factor.

$$\text{gender\_fac} \;=\; \begin{cases} 1.1, & \text{male}, \\ 1.0, & \text{female}. \end{cases}$$

Fat-percentage modifier.

$$\text{fat\_mod} \;=\; 1 \;+\; 0.001 \cdot \%\text{fat}.$$

Distance factor (walking/running only). For distance-sensitive activities (walking, treadmill running, running), let dist_{km} be the covered distance (km):

$$\text{dist\_fac} \;=\; \begin{cases} 1.35, & \text{dist}_{\text{km}} \ge 30, \\ 1.25, & 21.1 \le \text{dist}_{\text{km}} < 30, \\ 1.15, & 10 \le \text{dist}_{\text{km}} < 21.1, \\ 1.00, & \text{dist}_{\text{km}} < 10, \end{cases} \quad \text{else dist\_fac} = 1.00.$$

Final water-intake label. A 10% buffer is applied and values are clipped to a physiological range:

$$\text{Water\_Intake}_L \;=\; \min\Big(\max\big(\tfrac{C}{700}\cdot m_I\cdot\text{fat\_mod}\cdot\text{gender\_fac}\cdot\text{temp\_fac}\cdot\text{dist\_fac}\cdot 1.1,\ 0.1\big),\ 6.0\Big).$$

The session sweat-loss estimate is:

$$\text{sweat\_loss}_L = (r_0 \cdot m_I + \Delta_T + \Delta_W) \times d_h$$

Finally, the water-intake target is clipped to physiologically reasonable bounds:

$$\text{Water\_Intake}_L = \min\left(\max(\text{sweat\_loss}_L, 0.25), 2.00\right)$$

This generated a single Water_Intake_L column used as the supervised label. During deployment, the trained model predicts this value from user-provided exercise plan, physiological attributes, and temperature, ensuring that recommendations reflect estimated sweat losses under the given conditions.

The model was trained with a dataset that includes average heart rate, resting heart rate, and maximum heart rate during sessions, as well as body fat percentage. However, in the app, it was decided not to require these inputs, reducing some of the accuracy while maintaining the app's accessibility. This is aimed at beginner athletes who often don't invest in smartwatches initially due to the high cost.

The implemented system transforms the modular design into an operational smart hydration planning tool, integrating machine learning with an interactive web interface. Development followed the CRISP-DM methodology, progressing from model training to deployment in a production-ready format.

Model Training and Optimization

The system was trained on a combined dataset consisting of gym members' exercise datasets and personal exercise logs. All features were preprocessed using scikit-learn's ColumnTransformer to apply StandardScaler for numerical features and OneHotEncoder for categorical variables. The feature engineering steps generated derived attributes such as Body Mass Index (BMI) and Exertion Score.

A Stacking Regressor was implemented, combining:

- **Linear Regression** for baseline linear trends,
- **Random Forest Regressor** for capturing non-linear interactions,
- **XGBoost Regressor** for iterative error correction.

The meta-learner, a Linear Regression model, aggregated predictions from the base learners. Hyperparameters for Random Forest and XGBoost were optimized to minimize RMSE.

Pipeline Construction

A single machine learning pipeline integrated the preprocessing and prediction stages, ensuring that the deployment environment could process user input in the same way as training data, ensuring its integrity and ease of application deployment in cloud environments.

Deployment

The front-end was developed using Streamlit, chosen for its lightweight, user-friendly design, and real-time interactivity. The app allows users to input details about their planned workout, physiological data, and ambient temperature. Forecasts are generated upon submission and displayed on the screen.

System Performance

The final stacked model achieved an RMSE of 0.231 liters, a Mean Absolute Error (MAE) of 0.334 liters, and an R^2 score of 0.801, outperforming the other individual base models and the performance benchmark set during the design phase. The deployment was tested to ensure responsiveness and usability.

6 Evaluation

This section presents the evaluation of the Intelligent Hydration Planning system, focusing on the accuracy, stability, and interpretability of the final model, as well as the implications of these findings for practical application. The results are critically analyzed within the context of the research objectives and compared with related works identified in the literature review.

6.1 Model Performance

The final stacked ensemble model was compared against its base learners using Mean Absolute Error (MAE) and the coefficient of determination (R^2). Table 1 summarises the performance metrics on the held-out test set.

Table 1: Model performance comparison on the test set

Model	MAE (L)	R^2
Linear Regression	0.410	0.794
Lasso Regression	0.355	0.790
Ridge Regression	0.349	0.794
KNN Regressor	0.391	0.714
Random Forest Regressor	0.343	0.769
XGBoost Regressor	0.346	0.769
LightGBM Regressor	0.363	0.747
Keras (Neural Network)	0.371	0.789
Stacking Regressor	**0.332**	**0.802**

The stacked model outperformed all individual learners, achieving the lowest MAE across all tested models. The R^2 score of 0.802 indicates that the model explains roughly 80% of the variance in the target hydration volume, confirming its strong predictive performance.

6.2 Error Analysis

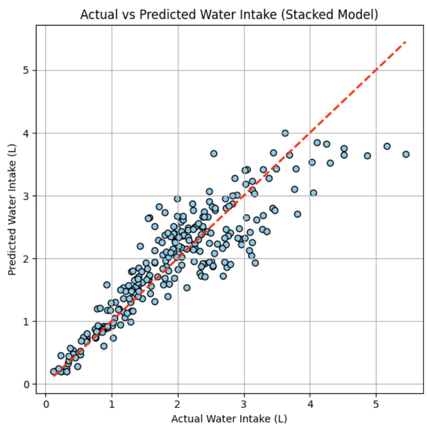

Figure 2: Residuals vs. Predicted Hydration Volume for the Stacking Regressor

Residual analysis was conducted to assess the distribution of bias and error. The residual versus predicted plot for the stacked model shows errors symmetrically distributed around zero, with no evidence of heteroscedasticity. This suggests that the model performs consistently at both low and high hydration needs, albeit with some outliers.

6.3 Model Explainability

To ensure transparency and build user confidence, SHAP (SHapley Additive ExPlanations) analysis was applied to the final model. Figure 3 shows that exercise duration and ambient temperature were the most influential factors, with positive contributions to predicted hydration volume.

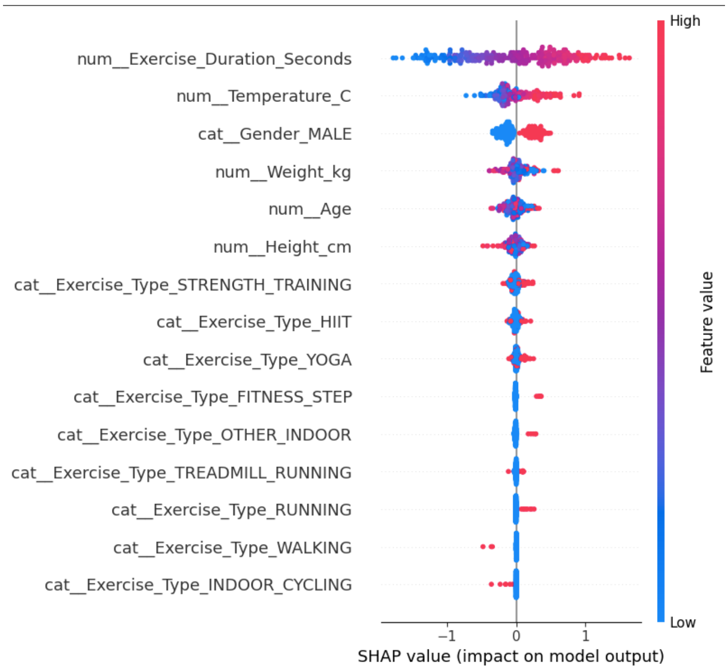

Figure 3: SHAP summary plot for the Stacking Regressor

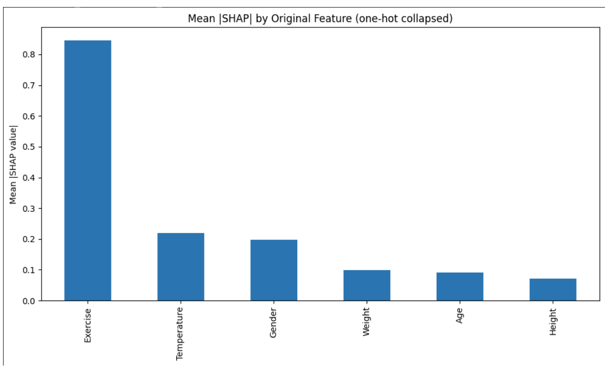

Figure 4: Mean absolute SHAP values by original feature (one-hot collapsed) for the stacked model. Exercise type is the dominant predictor, followed by temperature and gender, highlighting the importance of both activity characteristics and environmental conditions in determining hydration needs.

6.4 Comparison with Literature

The performance of the model aligns with the findings of Gray et al. (2023), who emphasised that reliable hydration guidance can be generated from a combination of user-provided workout and environmental parameters without the need for continuous biometric monitoring. The feature importance rankings also support the physiological models described by Casa et al. (2019), confirming that duration, temperature, and body mass are primary determinants of sweat loss.

6.5 Implications

From an academic perspective, these results demonstrate that ensemble learning can effectively integrate into the world of sports to improve hydration estimation accuracy. From a practitioner perspective, the system's reliance on readily available inputs makes it deployable in fitness and recreational sports contexts without expensive sensors or wearable devices.

6.6 Limitations and Potential Improvements

Although the system achieves good predictive performance, its accuracy is limited by the quality and diversity of the training data. Including historical user data (physiological and training) could improve prediction and performance, enhancing personalization. Including additional environmental variables, such as humidity and sun exposure, could further refine predictions. Furthermore, while SHAP analysis improves interpretability, real-time validation with diverse athlete populations would strengthen the model's generalizability.

7 Conclusion and Future Work

The research question addressed in this study was: *Can machine learning be used to generate accurate and personalized hydration recommendations for athletes, integrating details of planned training, physiological characteristics, and weather conditions, without the need for wearable sensors or continuous physiological monitoring?* The main objectives were: (1) to design and implement a predictive model using personal and publicly available training data, aggregated with weather information, (2) to identify the most influential factors affecting hydration needs, and (3) to implement the solution in an accessible and interactive web application.

To achieve these goals, a stacked regression model combining Random Forest, XGBoost, and Ridge Regression was trained on a dataset that integrated the gym members' exercise dataset, personal exercise logs, and temperature data. The feature engineering steps incorporated derived metrics such as BMI, pace, effort score, and distance sensitivity, while temperature was used to model heat-related hydration adjustments. The application was established via Streamlit, enabling real-time predictions from user-provided exercise, physiological, and climate data.

The solution successfully met the stated objectives. The final model achieved an MAE of 0.332 L and an R^2 score of 0.802, indicating that it explains approximately 80% of the variance in hydration needs. Residual analysis confirmed consistent performance across both low and high hydration demands, while SHAP analysis identified *Exercise Type*, *Temperature*, and *Gender* as the most influential predictors. These results align with sports science literature, validating the model's capacity to reflect physiological realities in hydration planning and the critical role of environmental conditions.

The implications of these findings are numerous. Academically, the project demonstrates that a non-invasive, data-driven approach can provide reliable, individualized hydration guidance by combining training, physiological, and climatic parameters, contributing to the literature on AI in sports science. From a practitioner perspective, the solution offers an accessible, evidence-based tool for athletes and coaches to optimize hydration strategies and reduce the risks associated with dehydration, especially in challenging environmental conditions.

The research is not without limitations. The dataset, while effective for model training, presented limited diversity, due to half of the records being personal. The dataset also had limitations due to its size and environmental conditions, and underrepresented exercise types. Furthermore, the model currently predicts a single water intake target for the planned activity, rather than dynamic adjustments during the session.

For future work, several meaningful avenues exist:

- **Integration with real-time data sources:** Incorporating optional wearable device inputs, such as heart rate and sweat rate, for dynamic adjustments to hydration targets.
- **Population-specific modelling:** Developing specialized models for different athletic populations (e.g., endurance runners, team sports players, youth athletes) to improve personalization.
- **Environmental adaptation:** Expanding the dataset to include more extreme climate conditions and seasonal variations, enabling better generalization in hot, humid, cold, or high-altitude environments.
- **Historical performance modelling:** Enhancing historical adaptation by integrating long-term performance trends and recovery data for more accurate, user-specific recommendations.
- **Commercial integration:** Embedding the prediction system into popular fitness tracking platforms or releasing it as a standalone mobile app to reach a wider audience.

In conclusion, this work demonstrates that machine learning can effectively provide personalized hydration guidance by combining training data, physiological attributes, and weather conditions without the need for expensive sensor systems. With increased dataset availability, real-time adaptability, and integration into commercial ecosystems, the solution has the potential to improve athletic performance, reduce injury risk, and enhance long-term recovery strategies.

References

Belabbaci, N. A., Anaadumba, R. and Alam, M. A. U. (2025). Recent advancements in wearable hydration-monitoring technologies: Scoping review of sensors, trends, and future directions, *JMIR mHealth and uHealth* **13**: e60569.

Benavides, N. R., Rutkey, H. E., DiDomenico, C. M., Wong, T., Martel-Foley, J., Yu, C.-H. and Kiapour, A. (2021). A wearable mobile-app controlled device for continuous monitoring of uv exposure and hydration levels, *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, pp. 1167–1172.

Casa, D. J., Cheuvront, S. N., Galloway, S. D. and Shirreffs, S. M. (2019). Fluid needs for training, competition, and recovery in track-and-field athletes, *International Journal of Sport Nutrition and Exercise Metabolism* **29**: 175–180.

Chodkowski, J. (2024). Role of nutrition and hydration in injury prevention and recovery: A review, *Journal of Education, Health and Sport* **71**: 56117.

Galekwa, R. M., Tshimula, J. M., Tajeuna, E. G. and Kyandoghere, K. (2024). A systematic review of machine learning in sports betting: Techniques, challenges, and future directions.
URL: *http://arxiv.org/abs/2410.21484*

Gray, M., Birkenfeld, J. S. and Butterworth, I. (2023). Noninvasive monitoring to detect dehydration: Are we there yet?, *Annual Review of Biomedical Engineering* **25**: 23–49.

Gupta, R. S., Wood, C. E., Engstrom, T., Pole, J. D. and Shrapnel, S. (2025). A systematic review of quantum machine learning for digital health, *npj Digital Medicine* **8**: 237.

Holland-Winkler, A. M. and Hamil, B. K. (2024). Hydration considerations to improve the physical performance and health of firefighters, *Journal of Functional Morphology and Kinesiology* **9**: 182.

Judge, L., Bellar, D., Popp, J., Craig, B., Schoeff, M., Hoover, D., Fox, B., Kistler, B. and Al-Nawaiseh, A. (2021). Hydration to maximize performance and recovery: Knowledge, attitudes, and behaviors among collegiate track and field throwers, *Journal of Human Kinetics* **79**: 111–122.

Kotowicz, M., Bieniak-Pentchev, M. and Koczkodaj, M. (2025). Artificial intelligence in medicine: Potential and application possibilities - comprehensive literature review, *Quality in Sport* **41**: 60222.

Leckey, C., van Dyk, N., Doherty, C., Lawlor, A. and Delahunt, E. (2025). Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis, *British Journal of Sports Medicine* **59**: 491–500.

Padulo, J. (2025). Sport and health science: interdisciplinary approaches to modern challenges, *British Medical Bulletin* **155**.

Papaoikonomou, G., Apergi, K. and Malisova, O. (2025). Children, adolescents and urine hydration indices—a systematic literature review on athletes and non-athletes, *Children* **12**: 171.

Pareek, A., Ro, D. H., Karlsson, J. and Martin, R. K. (2024). Machine learning/artificial intelligence in sports medicine: state of the art and future directions, *Journal of ISAKOS* **9**: 635–644.

Prova, N. N. I. (2024). Advanced machine learning techniques for predictive analysis of health insurance, *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, IEEE, pp. 1166–1170.

Ramírez., D. J. G. C., Islam, M. and Even, A. I. H. (2024). Machine learning applications in healthcare: Current trends and future prospects, *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023* **1**.

Richter, C., O'Reilly, M. and Delahunt, E. (2024). Machine learning in sports science: challenges and opportunities, *Sports Biomechanics* **23**: 961–967.

Sreeharsha, A., McHale, S., Nnamoko, N. and Pereira, E. (2024). Towards data-driven hydration monitoring: Insights from wearable sensors and advanced machine learning techniques, *Electronics* **13**: 4960.

Teo, Z. L., Jin, L., Liu, N., Li, S., Miao, D., Zhang, X., Ng, W. Y., Tan, T. F., Lee, D. M., Chua, K. J., Heng, J., Liu, Y., Goh, R. S. M. and Ting, D. S. W. (2024). Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture, *Cell Reports Medicine* **5**: 101419.

Włoch, T., Kwiatkowski, M., Koziński, T., Donderska, M., Bochniak, P., Kołacz, J., Oleś, P., Bryła, M., Bryła, Z. and Szpaczyńska-Kwiatkowska, M. (2025). Application and effectiveness of modern technologies in sports training - a literature review, *Quality in Sport* **42**: 60523.