Hackathon Project:

RPE prediction

Report

ANIZON Thibault

BRIVARY Guillaume

MARANDIN Tristan

MAUGER Mika

TROGNON Jean-Baptiste

WITKOWICZ Nathan

December 22th, 2023

Contents

Introduction 3
Dataset and Preprocessing 3
Random Forest Model Implementation 4
Results 5
FatigueNet Implementation 5
  Preprocessing 5
  Our deep learning model 6
Results 6
Discussion 7
Paths for Improvement 7
Conclusion and Future Work 8
References 9

Introduction

This report addresses the utilization of Data Science and Machine Learning techniques to enhance the performance of professional athletes, with a specific focus on professional football. The core challenge is the development of predictive models for Rate of Perceived Exertion (RPE), a crucial subjective metric in assessing training load and intensity. RPE reflects an athlete's perceived effort during physical activities, typically gauged on a scale of 1 to 10. The significance of this measure extends beyond sports performance, offering potential benefits in the medical field. The project's main objective is to explore the feasibility of predicting RPE using various data types specific to professional football players, including anthropometric data, GPS and accelerometer data during activities, heart rate, weather conditions, and historical RPE scores. This exploration is fundamental in advancing sports science and offers broader implications in health and fitness management.

The convergence of big data, AI, and ecological dynamics is reshaping our understanding and assessment of athletic performance, emphasizing the role of AI methodologies in sports science. This project's approach to RPE prediction using diverse athlete data mirrors these interdisciplinary methods [1]. Additionally, the use of Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM), which has shown high accuracy in analysing dynamic motion and classifying sports actions, underlines the potential of these techniques in our context for efficient sequential data processing and time-series analysis [2].

Dataset and Preprocessing

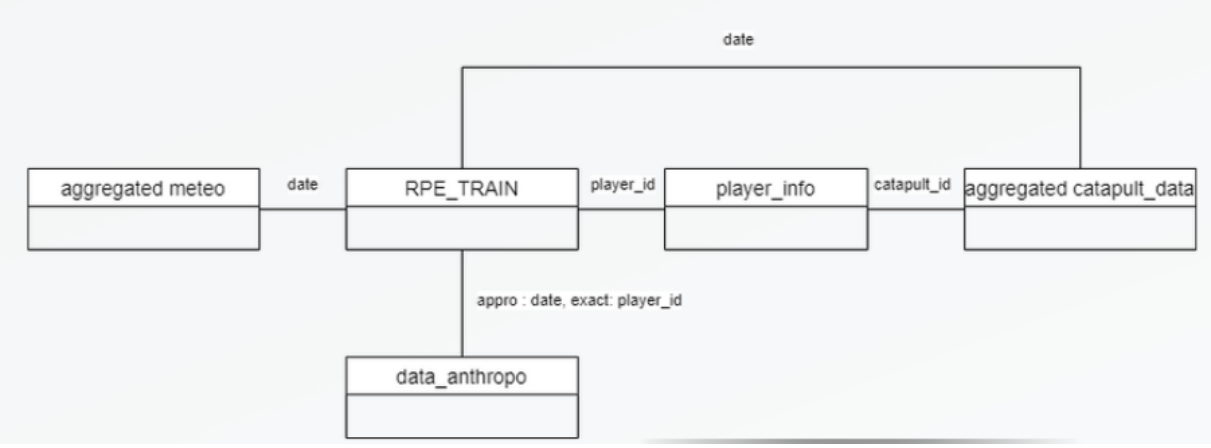
Our project faced several data-related challenges: the data was scattered across different tables requiring complex join operations, and it presented varying time scales necessitating thoughtful aggregations. To handle the issue of missing values, we employed different strategies such as filling in with mean values, zeros, or nearest values depending on the context.

The dataset itself is a rich amalgamation of multiple sources:

1. **Anthropomorphic Data**: Includes player-specific metrics such as age, BMI, fat and muscular mass percentiles, and height-weight ratios. Notably, BMI sometimes records as zero, requiring a threshold to be set for validity.
2. **Player Data**: Contains position codes, detailed position information, player identifiers, and catapult system IDs, essential for linking performance data to individual players.
3. **Catapult Data**: Provides detailed movement and physiological data such as GPS coordinates, velocity, acceleration, heart rate, and various performance load metrics.

4. **Meteorological Data**: Includes comprehensive weather-related data like temperature, humidity, wind speed and direction, and precipitation, crucial for contextualizing performance data.

This rich dataset, encompassing both individual player metrics and environmental factors, forms the foundation of our model development for predicting RPE in professional football players.

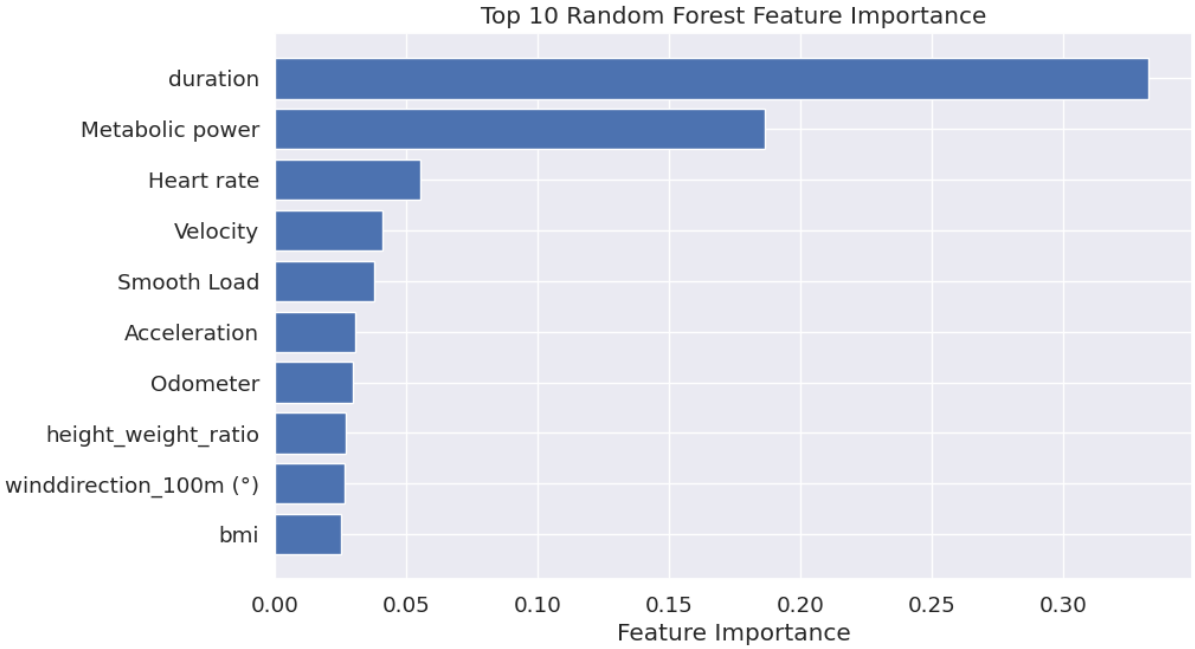


The database schema depicted in the graph outlines the structure and interconnections between the various datasets used in our analysis. The 'RPE_TRAIN' table, which records the Rate of Perceived Exertion, sits at the heart of the schema, linked to 'player_info' and 'aggregated_catapult_data' through the 'player_id' field. This indicates a relational model where the core assessment of training load is informed by detailed player metrics and activity data. The 'aggregated meteo' table is linked by date, suggesting that environmental factors are also considered in analyzing RPE. Each table serves a specific purpose, with 'data_anthropo' providing baseline player characteristics that do not frequently change. This structured approach ensures that analyses are comprehensive, accounting for both individual player differences and external factors affecting performance.

Random Forest Model Implementation

In the initial phase of our project, we implemented a Random Forest model [2]. This decision was driven by the model's ability to handle large and diverse datasets, its robustness against overfitting, and its interpretability. The Random Forest model was trained using a subset of the pre-processed dataset, focusing on key features that influence RPE, such as player-specific metrics and performance data. We optimized the model by tuning hyperparameters like the number of trees and depth of each tree. The model's performance was evaluated using Mean Absolute Error (MAE) against a validation set, providing a baseline for comparison with more complex models. This stage was crucial in understanding the data's underlying patterns and setting a benchmark for subsequent deep learning models.

Results



The Random Forest model's effectiveness in predicting Rate of Perceived Exertion (RPE) was evaluated, revealing the feature importance within the dataset. Duration was the most significant predictor, followed by metabolic power and heart rate, which aligns with physiological expectations of exertion. Other notable features included velocity, smooth load, acceleration, and the height-weight ratio. This model achieved a Mean Absolute Error (MAE) of 0.92, using a dataset with NaN values set to zero. The insights gained are invaluable for refining the model and improving predictive accuracy for RPE in professional athletes.

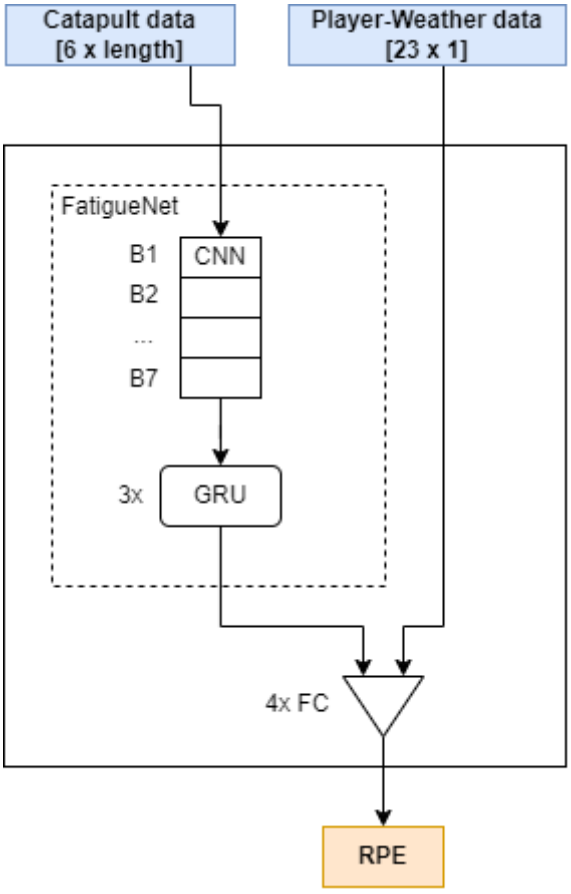
FatigueNet Implementation

Preprocessing

Our preprocessing pipeline for the varied dataset involved several steps to prepare the data for the FatigueNet model. We began by merging different data sources and normalizing the features to ensure consistent scale. The data was then split by 'player_id' to maintain individual player integrity and 'session_id' to segment the data into discernible sessions. Dates with inconsistencies were removed to maintain temporal coherence. Traditional cleaning methods were applied, along with a process to handle anomalous heart rate data. To accommodate the model's input requirements, we implemented padding for sequence data and re-normalized the dataset post-merging to ensure that the model received clean, structured, and normalized input. This comprehensive preprocessing was crucial for the subsequent modelling stages.

Our deep learning model

Following the Random Forest model, we progressed to implementing a model based on FatigueNet [4], a deep learning model, designed for time-series analysis, leverages deep convolutional layers and Gated Recurrent Units (GRUs) to process spatio-temporal data effectively. The architecture is specifically tuned to handle the complexity of the football data, capturing the intricate patterns in players' movements and physiological responses. The model was trained on the same dataset, emphasizing temporal dependencies crucial for predicting RPE. This advanced approach aimed to surpass the Random Forest model in accuracy and provide deeper insights into the factors affecting players' perceived exertion levels. FatigueNet's performance evaluation used the same MAE metric, allowing a direct comparison with the initial Random Forest model.



Results

The Random Forest model's MAE of 0.92 demonstrates its capability to discern complex patterns in the data, balancing well between underfitting and overfitting. This ensemble method's effectiveness in our context could stem from its nature of aggregating decisions from multiple decision trees, which tends to enhance the generalizability of the model. By capturing diverse relationships and interactions among the predictors, the model has shown a commendable level of accuracy. However, the real test of its

reliability as a forecasting tool for RPE will be its performance on new, unseen data, which could further validate its utility in practical settings.

The FatigueNet model's performance, with its nuanced architecture designed to process sequential and time-series data, brought forward the challenges of predicting RPE. The individual models' higher MAE of 1.7521 compared to the Random Forest might be hinting at overfitting issues, where the model is too closely attuned to the training data, lacking the ability to generalize. On the other hand, the global model's MAE of 3.1936 could be reflecting the difficulty of capturing the diverse physiological and performance responses of different players within a single predictive framework. These outcomes underscore the necessity for a balanced approach that can cater to individual player nuances without sacrificing the broader trends that apply across the team. Future work might explore hybrid models that combine individual player predictions with team-wide trends to optimize RPE prediction accuracy.

Discussion

Our analysis revealed that despite the limited data, the models delivered honourable scores. This performance suggests that even with constraints, valuable insights can be extracted. However, the deviation from trends observed in the literature prompts a re-evaluation of our approach, specifically considering the intricacies of our dataset and its potential idiosyncrasies compared to those commonly studied.

The relatively simple processing of meteorological data and player-specific information could be a contributing factor to the divergence from expected outcomes. In the literature, these variables often receive more detailed treatment, potentially through sophisticated feature engineering or the application of domain-specific transformations, which could account for their stronger predictive power. Enhancing the handling of these data types may provide a more nuanced understanding and lead to improved model performance.

Further discussions should focus on the depth of data processing methods applied to both meteorological and player-related features. Given their potential impact on athletic performance, a more granular analysis may reveal subtleties that refine model predictions. Optimizing data treatment could bridge the gap between our findings and established research, thereby enriching the predictive capability for RPE.

Paths for Improvement

To enhance the predictive power of our models, we can consider several improvement strategies. First, giving more relevance to the Catapult features could be crucial. These features, which include detailed player movement and biometrics, have the potential to deeply inform the model about physical exertion.

Second, incorporating a recurrent neural network (RNN) layer that takes into account data from previous sessions could offer a more dynamic temporal understanding of the athletes' fatigue and recovery cycles. This would allow the model to learn from sequences of activities, rather than from isolated sessions, providing a more holistic view of the athletes' condition over time.

Finally, the addition of cyclic variables is another avenue for improvement. Many features, such as time of day and seasonal aspects, are inherently cyclic and might have patterns that a standard neural network could miss. By transforming these into features that capture the cyclical nature—like using sine and cosine functions for time—we can enable the model to recognize and utilize the periodicity in the data, potentially improving the accuracy of RPE predictions.

Conclusion and Future Work

The experience garnered from this project extends far beyond technical proficiency in machine learning and data science. It encompasses the apprehension of complex problems in a real-world context, highlighting the importance of collaboration within a diverse, multi-skilled team. This collaboration not only facilitated effective team management but also fostered an environment conducive to learning and skill development. The project's challenges necessitated innovative problem-solving approaches, encouraging each team member to step outside their comfort zone and explore new methodologies and tools.

Looking forward, our roadmap includes a more granular approach to data collection, particularly focusing on session-based data which could provide deeper insights into the athletes' performance and well-being. This will be complemented by integrating the improvement strategies outlined earlier. We believe that these steps will not only refine our current models but also pave the way for more sophisticated analyses in the future.

In summary, this project has been a significant step in leveraging data science and machine learning for sports performance enhancement. It has provided a solid foundation and a clear direction for future research, which holds great promise for advancing not only sports science but also broader health and fitness applications. The knowledge and experience gained through this project will undoubtedly contribute to more effective strategies for athlete training, injury prevention, and overall well-being management.

References

[1] Araújo, D. *et al.* (2021) *Artificial Intelligence in sport performance analysis* [Preprint]. doi:10.4324/9781003163589.

[2] Fok, W.W., Chan, L.C. and Chen, C. (2018) 'Artificial Intelligence for sport actions and performance analysis using recurrent neural network (RNN) with long short-term memory (LSTM)', *Proceedings of the 2018 4th International Conference on Robotics and Artificial Intelligence* [Preprint]. doi:10.1145/3297097.3297115.

[3] Marqués-Jiménez, D. *et al.* (2023) 'A random forest approach to explore how situational variables affect perceived exertion of elite youth soccer players', *Psychology of Sport and Exercise*, 67, p. 102429. doi:10.1016/j.psychsport.2023.102429.

[4] Kim, J. *et al.* (2022) 'A deep learning approach for fatigue prediction in sports using GPS data and rate of perceived exertion', *IEEE Access*, 10, pp. 103056–103064. doi:10.1109/access.2022.3205112.