Journal of Physical Education and Sport ® (JPES), Vol. 24 (issue 8), Art 214, pp. 1927 - 1932, August 2024
online ISSN: 2247 - 806X; p-ISSN: 2247 – 8051; ISSN - L = 2247 - 8051 © JPES

Original Article

Athlete performance prediction using intelligent reporting and regression model analysis: A generative approach for training planning.

MOHAMED REBBOUJ[1], SAID LOTFI[2]
[1,2,]Multidisciplinary Laboratory In Education Sciences and Training Engeneering (LMSEIF). Sport Science Assessment and Physical Activity Didactic. Normal Higher School (ENS-C), Hassan II University of Casablanca, MOROCCO.

Published online: August 31, 2024
Accepted for publication : August 15, 2024
DOI:10.7752/jpes.2024.08214

Abstract:
This study explores the use of reported data and predictive analysis as a long-term generative approach for athlete training planning. The data collected from 607 higher education students (mean age = 16.86; Std = 1.22), includes measurements from physical tests and activity records. The dataset comprises 29 variables, which were profiled and feature-engineered to enhance predictive accuracy for training procedures. We leveraged Microsoft Azure Machine Learning to determine the significance of features on outcomes and utilized Power BI to visualize the impact of aggregate features on running distance. Initial findings indicates that the optimal age range for focusing training efforts is between 16 and 17 years old. This result is supported by a Spearman correlation coefficient of 0.42, stipulating a moderate positive relationship between the age group and the predicted performance outcomes based on key aggregate features. In particular, four key features significantly impact performance, while other variables have a minor influence. The study highlights the significance of these aggregate features in predicting training success. In conclusion, the study underscores the importance of a robust reporting process and the use of predictive analysis in developing training programs. It identifies four critical features that have a substantial impact on realized performance. While these four features are paramount, the study also acknowledges that other variables, though less influential, may still affect the outcomes if considered. This comprehensive approach to data collection and analysis provides a solid foundation for optimizing athlete training programs, ensuring that training efforts are both targeted and effective. The findings offer valuable insights for coaches and sports scientists aiming to enhance athletic performance through data-driven training strategies.
Keywords: Performance Optimization, Sports analytics, Data-Driven Training.

Introduction

Analysing the sport performance is a prevalent practice in the sport industry, where clubs and market bettors employ statistical methods and advanced predictive analysis to define the game or championship winner. Consequently, numerous models have been developed to enhance the prediction of a sport event outcomes (Hubáček et al., 2019; Maszczyk et al., 2014; McCabe & Trevathan, 2008; Zhang et al., 2022). These models rely on data from athletes, teams, or from external sources as betting odds, prognostics, and audience feedback. The collected information must be analysed in short time to facilitate decision-making through reporting, analysing, and predicting solutions.

Intelligent data analysis, a technology-driven process to derive a method, process, or action plan from data (Berthold & Hand, 2007). Is commonly used in reporting for decision-making process in various industry fields. To revolutionize the industry, a digital transformation is required for real-time reporting and prompt decision-making (Panchal et al., 2024). In the health sector Business Intelligence (BI) is used to obtain real-time data for visualization and decision-making toward medication trends (Frestel et al., 2023). Meanwhile in marketing, data analysis process generates reports to better position the product by segmenting customer choices. Business applications and software are often cloud-based solutions; hence, industries choose cloud solutions to pay the services they need when they need them. Microsoft Azure Machine Learning offers a more specific overview by providing data storage, computing power and predictive algorithms for industry-specific purpose, such as predicting next quarter's unit sales or a building's energy consumption (Shapi et al., 2021) . The health sector benefits significantly from this solution, using it to manage data about sick children in hospitals and facilitate the use of electronic health records (Guo et al., 2023). Additionally, the sports sector and related sciences use Azure Kinect for motion tracking and biomechanical analysis (Brambilla et al., 2023), and even to measure spatiotemporal parameters during walking (Guess et al., 2022).

Despite the extensive application of statistical methods and predictive analysis within the sports industry, a significant challenge persists in integrating diverse data sources and technologies to improve real-time reporting, predictive accuracy, and decision-making efficiency (Szymanski, 2020). The rapid analysis of

Corresponding Author: MOHAMED REBBOUJ, E-mail: Mohamed.rebbouj@enscasa.ma

data from athletes, teams, and external sources is crucial for determining game or championship outcomes. However, existing models frequently fail to deliver comprehensive, real-time insights essential for effective decision-making.

To effectively develop specific activities and target the necessary physical capacities to achieve desired performance outcomes (Eratlı Şirin & Şahin, 2020), it is essential to select appropriate candidates for each activity. Furthermore, tailored training plans for distinct groups of athletes enable a focused approach to various training methods, categorized by athlete group, age, or activity (Anderson & Hargreaves, 2016).

Our research seeks to address this challenge by leveraging advanced cloud-based solutions, specifically Microsoft Azure Machine Learning and Power BI, to develop an integrated framework for sports performance analysis. We utilize Azure's capabilities for data storage, computational power, and predictive algorithms to analyze physical performance data from athletes. Power BI is employed for dynamic reporting and visualization of the impact of aggregate features on performance metrics. This approach will facilitate the development of a generative solution for deferred decision-making, enabling coaches and physical education teachers to make informed decisions promptly based on real-time insights (Lath et al., 2021). By integrating these technologies, we aim to revolutionize the sports industry through a digital transformation that enhances training planning and performance prediction.

Materials and method

Data Collection

Data pertaining to 607 higher education students was meticulously collected through physical and sports tests during the academic year of 2022-2023. Subsequently, this data was classified into categorical and numerical forms to profile the dataset prior to conducting the analysis, as illustrated in the subsequent table.

Feature	Type	Mean	Std Dev	Variance	Skewness	Kurtosis
Sexe	Integer	1.51	0.5	0.25	-0.02	-2
Age	Integer	16.86	1.22	1.48	0.98	1.38
Poids	Decimal	60.4	9.84	96.88	0.55	0.26
Taille	Decimal	1.69	0.09	0.01	0.04	0.14
.....	.....	.....	.....	.....	.....	.....
Harvard T	Integer	75.05	20.53	421.62	0.12	-1.23
Détente V	Integer	67.42	33.14	1098.32	-0	-1.18
Détente H	Integer	196.39	74.3	5520.7	-0.002	-1.17

Table 1. Dataset profile of 33 columns and 607 rows.

The table provides a statistical summary of various features related to athlete performance, including sex, age, weight, height, and specific fitness test scores. Key points include the mean, standard deviation, variance, skewness, and kurtosis for each feature. The data reveals moderate variability in weight and height, significant variability in fitness test scores, and generally symmetrical distributions for most features. These insights are crucial for tailoring training programs and understanding athlete capabilities.

Statistical analysis

The dataset was imported into Power BI as a CSV file, and an automated report was generated to visualize the significant and correlated attributes. Simultaneously, the dataset was loaded into Azure ML to train a regression model, which helped identify the key features influencing sports performance (Rebbouj & Said, 2022). The data transformation process included steps such as preprocessing, feature engineering, application of scaling techniques, and the use of various algorithms to construct the model.

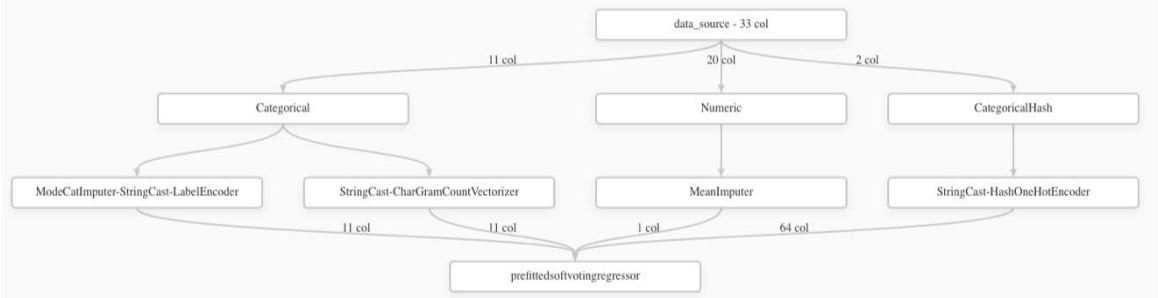


Figure 1. Data transformation process in Azure ML

The flowchart provides a structured approach to determining the appropriate analytical method based on the nature of the data source. It helps in selecting the right analysis technique, whether the data is categorical, numeric, or categorizable, thereby facilitating more accurate and efficient data analysis. This structured categorization is crucial for ensuring that the chosen analytical methods align with the data characteristics, ultimately enhancing the quality of insights derived from the analysis.

Results

The automated analysis conducted via Power BI underscored the significance of the heart rate (FC) feature, particularly in its correlation with the Body Mass Index (IMC) and the strength of the lower limb muscles (FMMI). In contrast, the aggregate feature importance derived from Azure ML identified the following variables as significant: weight (poids), the distance covered in 12 minutes (D-12min), the strength of the lower limb muscles (FMMI), and height (taille).

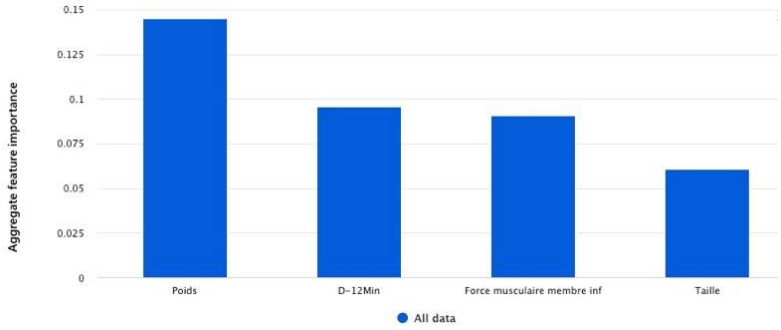
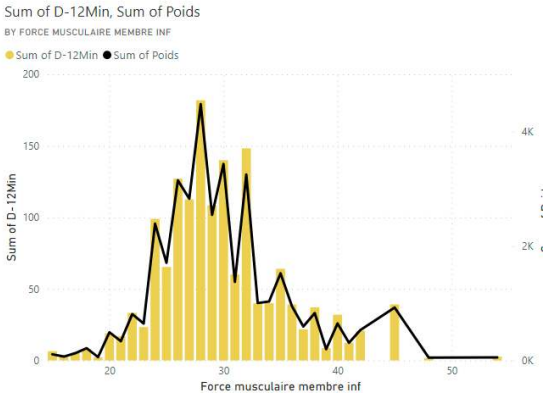


Figure 2. Aggregate feature importance for all data.

The graph helps identify which features are most critical for predicting outcomes, allowing for targeted improvements in training and performance strategies. By focusing on the most important features, such as weight and endurance, coaches and analysts can optimize training programs to enhance athlete performance.
The initial overview of the feature set will be employed within Power BI to enhance data visualization, encompassing the significance of each feature that has been empirically demonstrated to influence performance outcomes. Consequently, the generated visualizations depict a correlation between the Fat Mass and Muscle Index (FMMI), Weight, Height, and the performance achieved in a 12-minute distance run. This correlation provides valuable insights into the impact of these health metrics on physical performance.



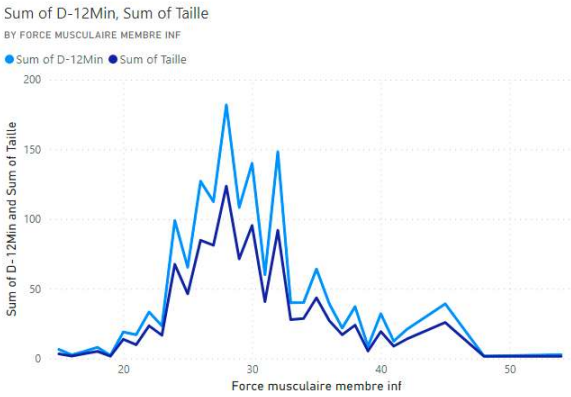


Figure 3 & 4. Impact of FMMI and the weight / Height on the Distance run in 12 min.

From the analysis of the two figures, it is inferred that a default decrease in muscle strength, as quantified by the Fat-Free Mass Index (FFMI), detrimentally influences the performance in achieving the maximum distance in the 12-minute Cooper test. This observation remains valid regardless of whether the athlete possesses the ideal weight or height required for the task. Importantly, age is considered a significant variable due to its substantial impact, as depicted in the subsequent figure. This consideration underscores the multifaceted nature of athletic performance, where factors such as muscle strength, body composition, and age collectively contribute to the outcome.

Figure 5. Predicted vs true chart of the regression model in azure ML.

In the context of training planning, the age bracket of 15.5 to 17 years is identified as the optimal range(Matušica & Peráček, 2023). This conclusion is drawn from the alignment of predicted and actual values as illustrated in the preceding chart. The model's metrics reveal a Spearman correlation of 0.42, signifying a substantial monotonic relationship. The model, as detailed in the data transformation process, employs a Voting Ensemble Algorithm. This approach underscores the importance of age in training planning and the effectiveness of ensemble methods in predictive modelling.

This graph is crucial for evaluating the accuracy of the predictive model. The closer the 'Average Predicted Value' line is to the 'Ideal' line, the more accurate the model. The bar chart provides additional context by showing the distribution of true values, helping to identify any patterns or biases in the model's predictions. Overall, this visualization aids in understanding the model's performance and highlights areas where improvements may be needed to enhance predictive accuracy.

Discussion

Upon the comprehensive analysis of data procured through Power BI and Microsoft Azure Machine Learning, tools we regard as deferred decision-making solutions over an extended period, we strategically reduced the sample size of student athletes from 607 to 416. This reduction, while significant, still ensures a robust participant pool for the validation of data post-training regimen completion. The selected 416 students have committed to undertaking the Cooper test, a 12-minute run, twice weekly. This commitment not only demonstrates their dedication(Ono et al., 2022) but also enriches our dataset. The inclusion of date and time variables in the dataset is a strategic move aimed at enhancing forecasting capabilities. This additional layer of data allows us to track progress over time and predict future performance trends, thereby refining the predictive accuracy of our model. This approach underscores the importance of comprehensive data collection and the effectiveness of machine learning tools in predictive modeling within the realm of athletic training.

The analysis of the data revealed several key insights. The bar graph titled "Aggregate Feature Importance" highlighted that "Poids" (Weight) is the most influential factor in the predictive model, followed by "D-12Min" (likely a 12-minute run test), "Force musculaire membre inf" (lower limb muscle strength), and "Taille" (Height). This information is crucial for tailoring training programs to enhance athlete performance(Tyshchenko et al., 2024).

Additionally, the "Predicted vs. True" graph provided a visual comparison between predicted values and actual true values. The 'Ideal' line represents a perfect prediction scenario, while the 'Average Predicted Value' line shows the actual performance of the predictive model. The slight fluctuations of the 'Average Predicted Value' line around the 'Ideal' line suggest that the model performs reasonably well, but there are some deviations. This visualization aids in understanding the model's performance and highlights areas where improvements may be needed to enhance predictive accuracy.

This methodology aims to establish a comprehensive framework for implementing a deferred information system within an educational environment (Krishnaveni & Meenakumari, 2010; Leidner & Jarvenpaa, 1993, 1995; Matušica & Peráček, 2023). The integration of Information and Communication Technology (ICT), particularly in the aftermath of the pandemic era, not only facilitates the digitization of the educational landscape but also enhances the efficacy of platforms in fostering teacher skill development(Ivanenko et al., 2020), streamlining analytical processes, and augmenting human-computer interaction. By leveraging advanced data analysis and machine learning tools, we can significantly improve the predictive modeling and decision-making processes in athletic training, ultimately leading to better performance outcomes for student athletes.

Conclusion

The amalgamation of two distinct solutions to generate critical information is a common practice in the industrial sector. This approach facilitates real-time reporting, expedites decision-making processes, and enhances financial outcomes by reducing time consumption. In contrast, the sports industry employs real-time decision-making during live events by analyzing data streams such as videos or images. This enables timely strategic modifications and provides crucial insights for success.

The objectives of our research were to optimize sports activities, enable students to achieve their peak performance, and ensure their participation in regional and national competitive sports through the use of advanced data analysis and machine learning tools.

Our analysis identified "Poids" (Weight) as the most influential factor, followed by "D-12Min" (12-minute run test), "Force musculaire membre inf" (lower limb muscle strength), and "Taille" (Height). This insight allows us to focus training programs on these key areas to optimize performance. The "Predicted vs. True" graph showed that our model performs reasonably well, with the 'Average Predicted Value' line closely following the 'Ideal' line. This indicates that our predictive model is effective in forecasting performance trends, which is crucial for optimizing training regimens.

By employing a deferred generative methodology, we focus on long-term decision-making in athletic training. This approach ensures that athletes are well-prepared for regional and national competitions. The integration of Information and Communication Technology (ICT) facilitates the digitization of the educational landscape, enhancing the efficacy of training platforms and fostering skill development.

To further enhance our approach, we recommend prioritizing training programs that target weight management, endurance (12-minute run test), and lower limb muscle strength, as these are the most influential factors identified in our analysis. We should continue to collect comprehensive data, including date and time variables, to improve forecasting capabilities and refine predictive models. Utilizing regression and time series forecasting techniques in Power BI and Azure Machine Learning will enhance predictive accuracy and provide deeper insights into performance trends. Encouraging athletes to maintain their commitment to regular training sessions, such as the Cooper test, will ensure continuous data enrichment and performance improvement.

References

Anderson, E., & Hargreaves, J. (2016). *Routledge handbook of sport, gender and sexuality*. Routledge London.

Berthold, M. R., & Hand, D. J. (2007). *Intelligent data analysis: an introduction*. Springer.

Brambilla, C., Marani, R., Romeo, L., Lavit Nicora, M., Storm, F. A., Reni, G., Malosio, M., D'Orazio, T., & Scano, A. (2023). Azure Kinect performance evaluation for human motion and upper limb biomechanical analysis. *Heliyon*, *9*(11), e21606. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e21606

Eratlı Şirin, Y., & Şahin, M. (2020). Investigation of factors affecting the achievement of university students with logistic regression analysis: School of physical education and sport example. *Sage Open*, *10*(1), 2158244020902082.

Frestel, J., Teoh, S. W. K., Broderick, C., Dao, A., & Sajogo, M. (2023). A health integrated platform for pharmacy clinical intervention data management and intelligent visual analytics and reporting. *Exploratory Research in Clinical and Social Pharmacy*, *12*, 100332. https://doi.org/https://doi.org/10.1016/j.rcsop.2023.100332

Guess, T. M., Bliss, R., Hall, J. B., & Kiselica, A. M. (2022). Comparison of Azure Kinect overground gait spatiotemporal parameters to marker based optical motion capture. *Gait & Posture*, *96*, 130–136. https://doi.org/https://doi.org/10.1016/j.gaitpost.2022.05.021

Guo, L. L., Calligan, M., Vettese, E., Cook, S., Gagnidze, G., Han, O., Inoue, J., Lemmon, J., Li, J., Roshdi, M., Sadovy, B., Wallace, S., & Sung, L. (2023). Development and validation of the SickKids Enterprise-wide Data in Azure Repository (SEDAR). *Heliyon*, *9*(11), e21586. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e21586

Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783–796.

Ivanenko, S., Tyshchenko, V., Pityn, M., Hlukhov, I., Drobot, K., Dyadechko, I., Zhuravlov, I., Omelianenko, H., & Sokolova, O. (2020). Analysis of the indicators of athletes at leading sports schools in swimming. *Journal of Physical Education and Sport*, *20*(4), 1721–1726.

Krishnaveni, R., & Meenakumari, J. (2010). *Usage of ICT for Information Administration in Higher education Institutions–A study*.

Lath, F., Koopmann, T., Faber, I., Baker, J., & Schorer, J. (2021). Focusing on the coach's eye; towards a working model of coach decision-making in talent selection. *Psychology of Sport and Exercise*, *56*, 102011.

Leidner, D. E., & Jarvenpaa, S. L. (1993). The information age confronts education: Case studies on electronic classrooms. *Information Systems Research*, *4*(1), 24–54.

Leidner, D. E., & Jarvenpaa, S. L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly*, 265–291.

Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. *Procedia-Social and Behavioral Sciences*, *117*, 482–487.

Matušica, M., & Peráček, P. (2023). The influence of relative age on the selection of players for elite events in junior football. *Journal of Physical Education and Sport*, *23*(3), 790–799.

McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. *Fifth International Conference on Information Technology: New Generations (Itng 2008)*, 1194–1197.

Ono, Y., Kaji, M., & Morita, T. (2022). A study of the worries that emerge in the career selection of Japanese student athletes. *Journal of Physical Education and Sport*, *22*(4), 1009–1017.

Panchal, G., Clegg, B., Koupaei, E. E., Masi, D., & Collis, I. (2024). Digital transformation and business intelligence for a SME: systems thinking action research using PrOH modelling. *Procedia Computer Science*, *232*, 1809–1818. https://doi.org/https://doi.org/10.1016/j.procs.2024.02.003

Rebbouj, M., & Said, L. (2022). Students' Physical Education Performance Analysis Using Regression Model in Machine Learning. *The International Conference of Advanced Computing and Informatics*, 682–692.

Shapi, M. K. M., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment*, *5*, 100037. https://doi.org/https://doi.org/10.1016/j.dibe.2020.100037

Szymanski, S. (2020). Sport analytics: Science or alchemy? *Kinesiology Review*, *9*(1), 57–63.

Tyshchenko, V., Tyshchenko, D., Andronov, V., Ivanenko, S., Adamchuk, V., Hlukhov, I., & Drobot, K. (2024). Comprehensive evaluation of efficiency to identify deficiencies in muscle activity in different modes in team sports. *Wiadomości Lekarskie Medical Advances*, *77*(2), 194–200.

Zhang, Q., Zhang, X., Hu, H., Li, C., Lin, Y., & Ma, R. (2022). Sports match prediction model for training and exercise using attention-based LSTM network. *Digital Communications and Networks*, *8*(4), 508–515.