Geosport for Society
ISSN 2393-1353, e-ISSN 2393-2353

vol. 22, no. 1/2025, pp. 1-9
DOI 10.30892/gss.2201-126

Machine Learning-Based Prediction of Algerian University Student Participation in Sports Activities

Mohamed Amine DAOUD[1*], Abdelkader BOUGUESSA[1], Kamel BENDDINE[2], Youcef DAHMANI[3]

[1] LRIAS Laboratory, Department of Computer Sciences, Ibn-Khaldoun University of Tiaret, Tiaret 14000, Algeria
[2] CEHM Laboratory, El-Bayedh University Center, BP 900, 32000, El-Bayedh, Algeria
[3] GEGI Laboratory, Department of Computer Sciences, Ibn-Khaldoun University of Tiaret, Tiaret 14000, Algeria

***Correspondence:** Mohamed Amine DAOUD; e-mail: mohamedamine.daoud@univ-tiaret.dz

Citation: Daoud, M.A., Bouguessa, A., Benddine, K., & Dahmani, Y. (2025). Machine Learning-Based Prediction of Algerian University Student Participation in Sports Activities. *Geosport for Society*, 22(1), 1-9. https://doi.org/10.30892/gss.2201-126

Article history: Received: 18.12.2024; Revised: 30.01.2025; Accepted: 08.02.2025, Available online: 13.02.2025

Abstract: Student participation in university sports is influenced by individual, social, cultural, and institutional factors. Despite the well-documented benefits of sports, barriers such as academic pressures and inadequate infrastructure hinder student involvement. This study employed a machine learning-based approach to predict sports participation among Algerian university students. Logistic regression and decision tree models were applied to analyze the dataset, focusing on key factors such as gender and athletic background. The models effectively predicted participation patterns, identifying gender and athletic background as significant determinants. Additionally, the analysis highlighted the most attractive sports disciplines for students, facilitating improved resource allocation and program design. The findings underscore the potential of machine learning to enhance university sports management. By providing actionable insights, this approach can guide the development of inclusive policies, fostering a dynamic and accessible sports ecosystem in Algerian universities.

Keywords: machine learning, prediction, sport, student participant

Introduction

Universities play a crucial role in fostering student socialization by fostering an environment that encourages exchanges and social interactions, particularly through sports activities. By offering opportunities to practice various sports, higher education institutions aim not only to improve students' physical health but also to foster essential skills for their social development, such as teamwork, discipline, and respect for others (Bailey et al., 2013, Stodolska et al., 2015). These activities, which are an integral part of the university experience, contribute to the development of social awareness by enabling students to become familiar with cooperation and community involvement.

University sports will ensure the alignment of athletic achievements to enhance the physical potential of younger generations through physical exercise (Wang, 2016) which is the primary and most effective means for this transformation. University sports hold a crucial place in students' academic and social journey, playing a vital role in their physical, mental, and social development (Eime, 2013). Within universities, sports are not merely leisure activities but a powerful tool for fostering social cohesion, personal discipline, and student well-being. In Algeria, where universities welcome a growing influx of students each year, sports participation remains limited to a few participants and disciplines. This situation highlights a significant challenge: despite access to sports facilities and the well-known benefits of physical activity, a low proportion of students engage in the sports activities offered.

This issue prompts us to investigate the reasons behind the limited participation and the factors that shape students' decisions to participate in or refrain from university sports. Therefore, it is relevant to conduct an in-depth study to analyze students' sports habits, understand perceived motivations and obstacles, and ultimately predict students' participation profiles (Hoffmann, 2009). By identifying these factors, this research aims to contribute to developing tailored strategies that encourage broader sports participation within Algerian universities, thereby enabling students to fully benefit from the advantages of physical activity and university life (Andrews, 2005, Brady, 2005).

The supervised or unsupervised learning (Bishop, 2006) used in this context will enable us to predict student participation based on many variables such as gender, age, sporting background, preferences for some disciplines, and even perceptions of the benefits of practicing sports. Based on complex data analysis (Chawla, 2024), this approach will open new perspectives for identifying the factors influencing student engagement in university sports. As a result, it will provide valuable insights to design targeted and practical strategies to encourage better and more inclusive participation in university sports activities.

Influence of Social Group: Social surroundings, including friends and family, play an essential role in a student's decision to participate in sports activities (Dishman et al., 2005). Additionally, cultural norms and social expectations of each country or institution often influence participation. Studies have shown that men participate more frequently in university sports activities than women, although this difference tends to decrease in specific contexts (Browne et al., 2013). However, factors such as the perception of social acceptance of women’s sports and the availability of resources can influence female students' engagement. Access to sports facilities and their quality directly impact student participation. Studies have highlighted that quality and easily accessible facilities encourage more students to engage in sports (Eime et al., 2013). The existence of diverse sports programs and the involvement of universities in promoting sports are also key factors. Some universities offer flexible schedules, sports clubs, and competitive events to encourage participation (Trost et al., 2002).

Numerous studies in the social sciences, physical education, and psychology have examined university sports participation (Zhou, 2024). These studies have mainly sought to understand the factors influencing students' engagement in university sports activities and the consequences of this participation on their personal, social, and academic development.

Materials and methods

Data description

This dataset contains comprehensive information on university sports programs across various institutions in Algeria (Algerian, 2024). It includes data on student enrollment and sports participation categorized by gender and sport. The dataset can be used to analyze trends and gender disparities in university sports.

This dataset contains comprehensive information about university-level sports programs across institutions in Algeria, capturing student enrollment and sports participation by gender (Kotsiantis, 2004). Each row represents a unique record for a specific institution in a particular year, detailing the demographic of sports programs. This dataset can analyze trends in university sports and evaluate gender disparities in participation within Algerian universities' sports programs. The attributes are organized into several categories (Table 1).

The dataset provides a comprehensive overview of various aspects of university sports participation and enrollment. Firstly, institution information includes unique identifiers such as united and institution_name to distinguish each university. Additionally, location details such as (city_txt, state_cd, and zip_text) provide the geographical context of each institution. Secondly, classification data categorizes universities based on attributes such as size, focus, or sports participation levels, as represented by (classification_code, classification_name, and classification_other). The dataset also includes sector information (sector_cd and sector_name), which specifies whether the institution is public or private. Thirdly, enrollment data offers gender-specific enrollment figures. These include male (ef_male_count), female (ef_female_count), and total enrollment counts (ef_total_count) for each institution. Finally, the dataset captures detailed sports participation information. Sports program details are identified using sports_code and sports, indicating the types of sports offered. Gender-specific participation is documented through partic_men, partic_women, partic_coed_men, and partic_coed_women, capturing male, female, and mixed-gender team participants. Furthermore, total participation counts for men and women are summarized in sum_partic_men and sum_partic_women across all sports programs.

This structure provides a rich dataset for analyzing patterns in university enrollment and sports participation.

Table 1. Algerian University Sports Dataset

Attributes	Description
year:	The academic or calendar year during which the data was collected.
unitid:	A unique identifier for each institution, helping to distinguish between different universities.
institution_name:	The name of the educational institution or university.
city_txt:	The city where the institution is located.
state_cd:	The code representing the region or state within Algeria.
zip_text:	The postal or zip code of the institution's location.
classification_code:	A numeric or alphanumeric code representing the type or classification of the institution (e.g., by size, research focus, or sports level).
classification_name:	The name associated with the classification code, providing a more descriptive label for the institution type.
ef_male_count:	The number of enrolled male students in the institution.
ef_female_count:	The number of enrolled female students in the institution.

ef_total_count:	The total number of enrolled students, combining both male and female counts.
sector_cd:	A code indicating the institution's sector, such as public or private.
sector_name:	The name of the institution's sector.
partic_women:	The number of female participants in the sport or sports program
partic_coed_men:	The number of male participants in co-ed (mixed-gender) sports.
partic_coed_women:	The number of female participants in co-ed sports.
sports:	The specific sport or activity (e.g., football, basketball) for which the data is being recorded.

Machine Learning Algorithms

Logistic regression is particularly well-suited for binary and multiclass classification. It is an excellent choice for predicting participation, whether as participation versus non-participation or involvement in a specific discipline. It provides probabilities associated with each class, enabling the assessment of the likelihood that a student will participate in a given sports discipline. Its simplicity and ability to avoid overfitting make logistic regression a high-performing model, especially in contexts where explanatory variables directly influence the probability of participation, such as sports preferences, available time, or the desired level of competition (Das, 2024).

Decision trees will analyze the data by splitting features to classify students according to their participation probability in each sports discipline. This model also identifies the most critical variables for classification, making it easy to visualize the factors most influencing participation. Once trained, the decision tree can be applied to new data to predict a student's involvement in a specific discipline based on their characteristics and preferences (Song et al., 2015), (Schidler et al., 2024).

Methodology steps

Data preprocessing

In this phase, handling missing values is crucial to maintaining the integrity of the dataset and ensuring that analyses and models are accurate. Here are some common approaches for handling missing values, along with guidelines for deciding which approach to use:

Mean/Median Imputation: This method involves replacing missing values with the mean (or median) of the non-missing values in that column.

Label Encoding: Label encoding assigns a unique integer to each category of a categorical feature.

Model validation

Validating the model is essential in predictive modeling to assess its accuracy and reliability in real-world applications. For predicting student participation in sports, model validation ensures that the chosen algorithms generalize well to new data beyond the specific samples used for training. In this study, several validation methods were applied to verify the predictive models' performance and minimize overfitting or underfitting.

Train-Test Split: A basic yet practical approach to validation is splitting the dataset into two parts: a training set, used to fit the model, and a test set, used to evaluate its performance. This method allows for a straightforward assessment of how well the model can predict unseen data. Typically, an 80-20 split is used to ensure that the model has sufficient samples for training while leaving enough data for robust testing.

Evaluation Metrics: To measure the predictive accuracy, several metrics were used (Rainio, et al 2024), including:

Accuracy: The proportion of correctly predicted instances out of all predictions made. Accuracy is useful for an overall sense of correctness.

Precision: Precision assesses the accuracy of positive predictions (i.e., the proportion of true positive predictions among all predicted positives),

Recall: Recall measures the ability of the model to capture all relevant positive cases. These metrics are beneficial for understanding the model's performance on minority classes, such as groups of students with lower participation rates.

F1-Score: This metric combines precision and recall into a single score, offering a balanced view of the model's performance. It is beneficial in scenarios with imbalanced participation data across sports or student demographics.

Area Under the ROC Curve (AUC-ROC): This metric evaluates the model's ability to discriminate between classes across all decision thresholds, making it a robust choice for binary and multi-class classification tasks in sports participation prediction (Chang, 2024).

Results and discussions

Logistic Regression

Figure 1 illustrates the Receiver Operating Characteristic (ROC) curve for multiclass classification, demonstrating the performance of a logistic regression model in predicting student participation across various sports disciplines. Each colored curve corresponds to a specific class, representing the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for different sports disciplines. The diagonal gray line serves as a reference, indicating the performance of a random model; curves closer to this line suggest limited predictive ability.

The model's performance is further quantified using the Area Under the Curve (AUC) values, which measure the model's ability to distinguish between classes. In this case, the AUC values, ranging from 0.48 to 0.55, reveal a performance close to random guessing. These results highlight the model's difficulty in accurately classifying student participation for individual sports disciplines. The low AUC values raise concerns about the model's effectiveness, which could stem from various factors. Possible reasons include insufficiently discriminative data for each sports discipline, an unsuitable model choice, or suboptimal feature representation. Such limitations suggest that the model may require refinement or additional data preprocessing to improve its predictive capabilities.

The axes of the ROC curve provide further insight into the model's performance. The x-axis represents the False Positive Rate, capturing the proportion of incorrect positive predictions, while the y-axis represents the True Positive Rate, indicating the proportion of correct positive predictions. The proximity of the curves to the diagonal line confirms the model's inability to add significant value beyond random prediction.

In conclusion, the logistic regression model demonstrates limited effectiveness in predicting student participation across sports disciplines. The results underscore the need for further investigation to enhance the model's ability to distinguish between classes and improve its overall performance.

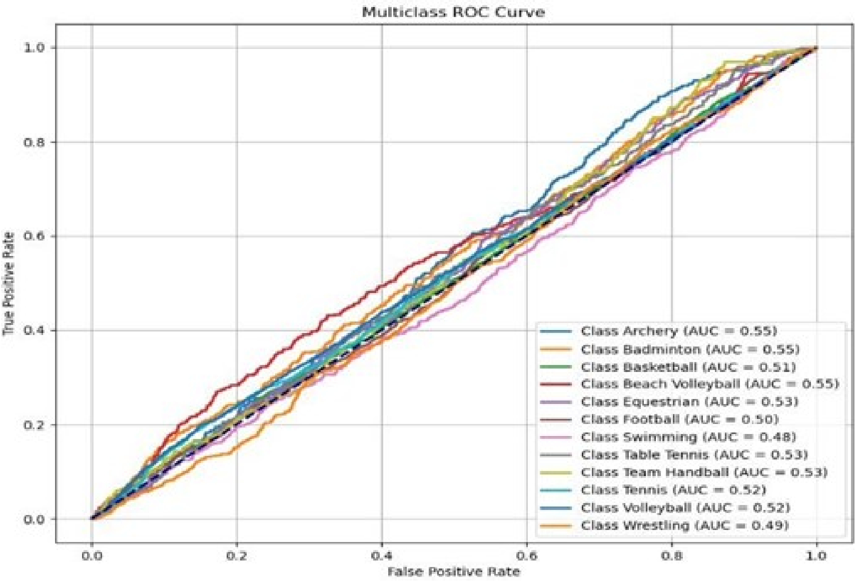

Figure 1. Logistic Regression ROC

Table 2 presents the confusion matrix for the model's performance for Logistic Regression model across multiple classes. The confusion matrix provides a detailed view of how well the model classifies instances into the correct categories. Each row in the matrix corresponds to an actual class, while each column represents a predicted class.

Table 2. Regression Logistic Confusion Matrix

Classes	Precision	Recall	F1-score	Support
Archery	0.00	0.00	0.00	288
Badminton	0.00	0.00	0.00	337
Basketball	0.21	1.00	0.35	2006
Beach Volley	0.00	0.00	0.00	394
Equestrian	0.00	0.00	0.00	381
Football	0.00	0.00	0.00	1049
Swimming	0.20	0.01	0.03	562
Table Tennis	0.00	0.00	0.00	327
Handball	0.00	0.00	0.00	298
Tennis	0.00	0.00	0.00	1297
Volleyball	0.00	0.00	0.00	1757
Wrestling	0.00	0.00	0.00	648

Decision Tree

The ROC curve for the decision tree model (Schidler, 2024), as shown in Figure 2, evaluates the performance of multiclass classification, likely representing various sports disciplines. The analysis of AUC (Area Under the Curve) values provides insight into the model's effectiveness across different classes.

The AUC values reveal significant variation across classes, ranging from 0.32 to 0.80. Notably, Class 11 achieves the highest AUC value of 0.80, indicating the model's strong performance in distinguishing this class from others. Conversely, classes such as Class 9 and Class 0, with AUC values of 0.32 and 0.35 respectively, demonstrate the model's struggle to differentiate these categories, performing only marginally better than random guessing. This variability in AUC scores highlights the model's differing levels of accuracy between classes. For instance, the higher performance in certain classes, like Class 11, suggests the presence of more distinct or discriminative features in the dataset. On the other hand, lower-performing classes may overlap significantly with other categories, posing challenges for the model.

The shapes of the ROC curves further reflect classification quality. Curves that approach the top-left corner signify better performance, yet in this case, only a few classes achieve this benchmark. Many curves closely align with the diagonal, indicating poor classification performance for those classes.

To enhance the model's predictive accuracy, several improvements could be considered. Fine-tuning the model, performing feature engineering, or employing more advanced classification algorithms might address current limitations. Additionally, obtaining more representative or distinct data for underperforming classes could help the model better differentiate between categories.

In conclusion, while the model demonstrates moderate success in predicting participation for certain classes, such as Class 11, its overall performance remains suboptimal, particularly for classes with low AUC values. Future enhancements could significantly improve its utility in this multiclass classification task

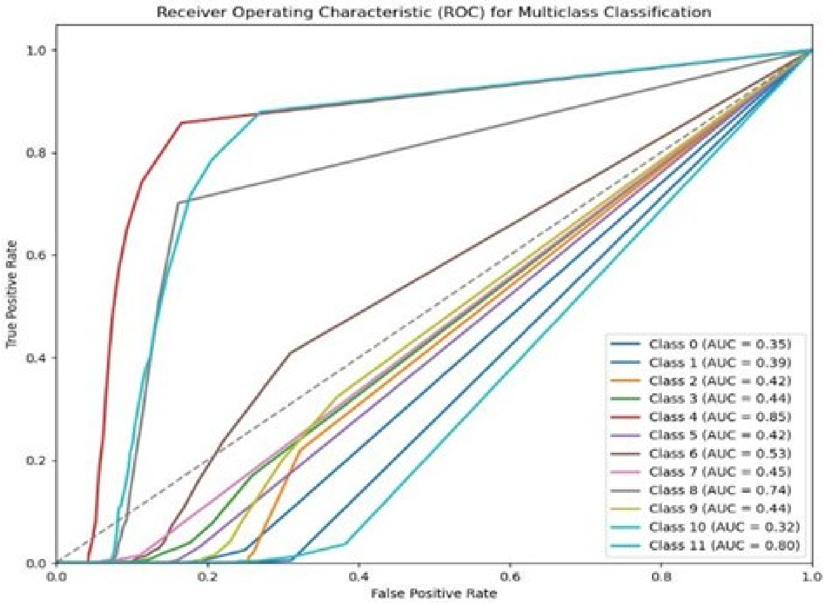

Figure 2. Decision tree ROC

Table 3 presents the confusion matrix for the model's performance for Decision Tree model across multiple classes. The confusion matrix provides a detailed view of how well the model classifies instances into the correct categories. Each row in the matrix corresponds to an actual class, while each column represents a predicted class.

Table 3. Confusion Matrix of Decision Tree

Classes	Precision	Recall	F1-score	Support
Archery	0.96	0.98	0.97	288
Badminton	0.99	0.99	0.99	337
Basketball	1.00	1.00	1.00	2006
Beach Volley	1.00	0.99	0.99	394
Equestrian	1.00	1.00	1.00	381
Football	1.00	1.00	1.00	1049
Swimming	1.00	1.00	1.00	562
Table Tennis	1.00	1.00	1.00	327
Handball	0.85	0.97	0.91	298
Tennis	0.99	0.96	0.97	1297
Volleyball	0.99	0.99	0.99	1757
Wrestling	0.97	0.97	0.97	648

Conclusions

In the context of promoting university sports activities in Algeria, this study proposed a machine learning-based approach to predict student participation in various sports disciplines. The primary objective was to identify the key factors influencing student engagement in sports, providing university decision-makers with decision-support tools to optimize sports policies and encourage more inclusive participation.

The machine learning models utilized, such as logistic regression and decision trees, demonstrated their effectiveness in predicting participants based on various demographic, academic, and sports-related characteristics. These results highlighted the most attractive disciplines and identified student groups most likely to engage in sports activities.

The proposed method offered a comprehensive understanding of the sports preferences of Algerian students, contributing to better resource planning and tailored sports programs. Moreover, this approach represents a significant advancement for the university sports community by supporting data-driven decision-making and fostering the development of a dynamic and inclusive sports ecosystem within Algerian universities.

The findings emphasize the potential of machine learning as a strategic tool for analyzing student behaviors and improving university sports management practices. Future research could explore more diverse datasets and implement more complex models to further enhance the impact of this approach on the student sports community.

Author contributions: Conceptualization, M.A.D.; methodology, Y.D.; formal analysis, M.A.D.; investigation, K.B.; writing - original draft preparation, A.B.; writing - review and editing, M.A.D. and A.B.; supervision, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this paper may be obtained on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

Algerian University Sports Federation. www.fasu-algeria.org. (Accessed on: 20.11.2024)

Andrews, J., Williamson, K., & Miller, R. (2005). Sport participation and its effect on academic performance and social integration in university students. *Journal of College Student Development*, 46(2), 179-194.

Bailey, R., Hillman, C., Arent, S., & Petitpas, A. (2013). Physical activity: An underestimated investment in human capital? *Journal of Physical Activity and Health*, 10(3), 289-308.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer, New York, USA.

Brady, L.L., Cagle, J.D., & Martin, J. (2017). The relationship between physical activity and perceived health benefits among university students. *Journal of Physical Activity and Health*, 14(5), 372-379.

Chang, P.W., & Newman, T.B. (2024). Receiver Operating Characteristic (ROC) Curves: The Basics and Beyond. *Hospital Pediatrics*, 14(7), e330-e334.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357

Das, A. (2024). Logistic regression. In Encyclopedia of Quality of Life and Well-Being Research (pp. 3985-3986). Cham: Springer International Publishing.

Dishman, R.K., Sallis, J.F., & Orenstein, D.R. (2005). The determinants of physical activity and exercise. *Public Health Reports*, 100(2), 158-171.

Eime, R.M., Sawyer, N.A., Harvey, J.T., & Casey, M.M. (2013). Participation in sport and physical activity: Associations with socio-demographic factors. *BMC Public Health*, 13, 191.

Eime, R.M., Young, J.A., Harvey, J.T., Charity, M.J., & Payne, W.R. (2013). A systematic review of the psychological and social benefits of participation in sport for children and adolescents: Informing development of a conceptual model of health through sport. *International Journal of Behavioral Nutrition and Physical Activity*, 10(1), 98.

Hoffmann, S., & Machida, S. (2009). The relationship between university sport participation and social and psychological outcomes. *European Journal of Sport Science*, 9(4), 331-339.

Kotsiantis, S.B., & Pintelas, P.E. (2004). Predicting Students' Performance with Machine Learning Techniques. *Proceedings of the 6th Hellenic Conference on AI,* 245-258.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1), 6086.

Schidler, A., & Szeider, S. (2024). SAT-based decision tree learning for large data sets. *Journal of Artificial Intelligence Research*, 80, 875-918.

Stodolska, M., & Floyd, M.F. (2015). Benefits of physical activity in higher education: Enhancing health and fostering essential social skills. *Journal of Higher Education and Physical Education Studies*, 12(3), 150–165.

Trost, S.G., Blair, S.N., & Moore, R. (2002). Physical activity and public health: A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *Journal of the American Medical Association*, 273(5), 402-407.

Wang, Y., Zhang, H., & Li, Q. (2014). Sport participation and academic achievement among university students in China. *Journal of Physical Education and Sport*, 14(2), 141-147.

Zhang, T., & Tsang, I.W. (2007). A Survey on Machine Learning Techniques and Their Application to Data Mining. *Journal of Data Mining and Knowledge Discovery*, 15(4), 247-270.

Zhou, Y. (2024). Sports College Education Under the Background of the Development of Sports Undergraduate Education. *Revista de Psicología del Deporte*, 33(2), 139-147.