STUDIA UBB EDUCATIO ARTIS GYMN., LXIX, 4, 2024, pp. 93-104
(RECOMMENDED CITATION)
DOI:10.24193/subbeag.69(4).32

# MACHINE LEARNING - BASED PREDICTION OF ALGERIAN UNIVERSITY STUDENT PARTICIPATION IN SPORTS ACTIVITIES

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

*Article history: Received: 2024 November 07; Revised 2025 January 20; Accepted 2025 January 22; Available online: 2025 February 10*; *Available print: 2025 February 28*


**ABSTRACT.** Student participation in university sports is influenced by individual, social, cultural, and institutional factors. Despite the recognized benefits of sports, many students face barriers such as academic pressures and inadequate infrastructure. This study proposed a machine learning-based approach to predict sports participation among Algerian university students, focusing on identifying key factors like gender and athletic background to guide inclusive sports policies. Using models like logistic regression and decision trees, the study effectively predicted participation patterns and highlighted the most attractive sports disciplines, enabling better resource planning and tailored programs. This approach offers valuable insights for fostering a dynamic, inclusive sports ecosystem and emphasizes the potential of machine learning to enhance university sports management.

***Keywords***: *Sport; University; Prediction; Activity; Algeria*

## INTRODUCTION

Universities play a central role in the socialization of students by creating an environment where exchanges and social interactions are encouraged, particularly through sports activities. By offering opportunities to practice

[1] LRIAS Lab, Dept. of Computer Sciences, Ibn-Khaldoun University of Tiaret, Algeria
[2] CEHM Lab, El-Bayedh University Center, Algeria
[*] Corresponding author: k.beneddine@cu-elbayadh.dz

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, 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.

Universities sports will ensure the alignment of athletic achievements with the goal of enhancing the physical potential of younger generations through physical exercise, which is the primary and most effective means for this transformation. University sports hold a crucial place in the academic and social journey of students, playing a vital role in their physical, mental, and social development (Eime, 2013). Within universities, sports are not merely leisure activities; they are 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 small number of 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 leads us to explore the reasons behind this limited participation and the factors that influence students' decisions to become involved, or not, in university sports. It is therefore relevant to conduct an in-depth study to analyze students' sports habits, understand perceived motivations and obstacles, and ultimately predict students' participation profiles. By identifying these factors, this research aims to contribute to the development of tailored strategies that encourage broader and more inclusive sports participation within Algerian universities, thereby enabling students to fully benefit from the advantages of physical activity and university life.

Faced with the issue of low student participation in university sports activities, it becomes essential to understand the various factors that influence their choices. Traditional approaches to data collection and analysis are not always sufficient to grasp the complexity of these behaviors, which are often marked by multiple and interconnected elements. This is why machine learning techniques, an advanced branch of artificial intelligence, are particularly well-suited for this study. These techniques allow for the analysis of large amounts of data to extract hidden patterns, revealing trends and correlations that are not immediately visible.

The use of supervised or unsupervised learning in this context will enable us to predict student participation based on variables such as gender, age, sporting background, preferences for certain disciplines, and even perceptions

of the benefits associated with practicing sports. This approach, based on the analysis of complex data, 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 effective strategies aimed at encouraging greater and more inclusive participation in sports activities within universities.

## METHODOLOGY

### Data Description (Algerian University Sports Dataset)

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

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

1. **Institution Information**:
   - **Institution identifiers** (such as unit id and institution_name) provide unique identification for each university.
   - **Location details** (city_txt, state_cd, and zip_text) indicate the geographical context of the institution.
2. **Classification Data**:
   - **Institution classification** (classification_code, classification_name, classification_other) categorizes the university by size, focus, or sports participation level.
   - **Sector information** (sector_cd, sector_name) indicates whether the institution is public or private.
3. **Enrollment Data**:
   - **Gender-based enrollment counts** (ef_male_count, ef_female_count, and ef_total_count) provide the total male, female, and combined enrollment figures for each institution.

4. **Sports Participation**:

- *Sports program codes and descriptions* (sports code and sports) specify the types of sports offered.
- **Gender-specific and co-ed participation counts** (partic_men, partic_women, partic_coed_men, partic_coed_women) capture the number of male and female participants in each sport, including mixed-gender teams.
- **Total participation counts** (sum_partic_men, sum_partic_women) summarize the male and female participants across all sports programs.

**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. |

## Justification of Machine Learning Algorithms

- Logistic regression

Logistic regression is particularly well-suited for binary and multiclass classification, making it an excellent choice for predicting participation, whether as participation versus non-participation or participation 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

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

**Proposed approach**

**a) Data Preprocessing**

**Data Cleaning**: In the data cleaning phase, handling missing values is crucial to maintaining the dataset's integrity and ensuring that analyses and models built on it 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.

**b) Model Validation**

In predictive modeling, validating the model is essential 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 performance of the predictive models and to minimize overfitting or underfitting.

- **Train-Test Split**:

A basic yet effective 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 make predictions on unseen data. Typically, an 80-20 split is used to ensure that the model has a sufficient number of 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 particularly useful 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 particularly helpful in scenarios where there is an imbalance in 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

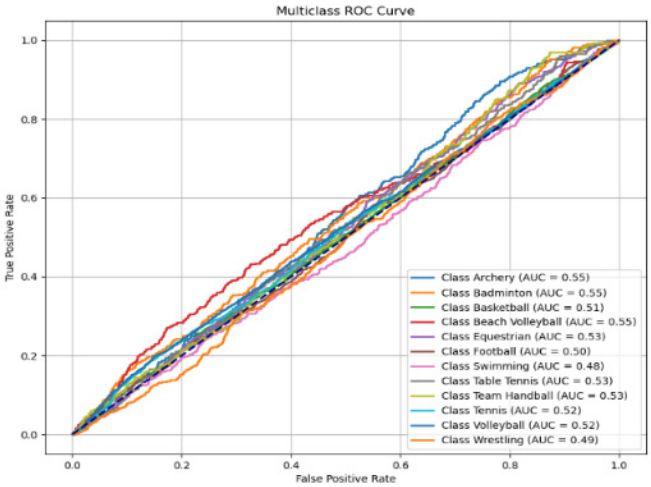


**Figure 1**. Logistic Regression ROC

Improving performance may require adjusting the model, adding more discriminative features, or using more sophisticated algorithms for this multiclass classification task.

**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

Figure 3 shows a Decision tree ROC curve for multiclass classification, evaluating the performance of a model across multiple classes (likely representing different sports disciplines). An interpretation of the results based on the AUC values is presented below:

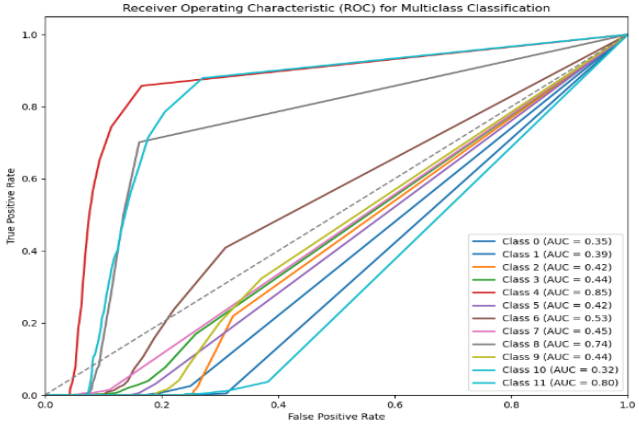


**Figure 2**. Decision tree ROC

**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 |

## DISCUSSION

### Logistic Regression

Figure 2 shows a Logistic Regression ROC (Receiver Operating Characteristic) curve for multiclass classification of student participation across various sports disciplines. The interpretation of the results is below:

- **Individual ROC Curves for Each Class**:

Each colored curve represents a specific class (a sports discipline). The curves show the relationship between the True Positive Rate and the False Positive Rate for different classes within the multiclass classification framework.

- **AUC (Area Under the Curve) Values**:

The AUC measures the model's ability to distinguish between classes. The AUC values indicated in the legend for each class range from 0.48 to 0.55. An AUC close to 0.5 means that the model performs similarly to random guessing for that class. In Figure 2, the AUC values are close to 0.5 for all classes, which suggests that the model struggles to correctly predict participation for each sports discipline.

- **Performance Interpretation**:
  - A higher AUC (close to 1) would indicate good classification performance, but here, the low AUC values suggest that the model is not effective in accurately distinguishing between different classes.

- The low AUC values for all classes could be due to several factors, such as a lack of discriminative data for each sport, an unsuitable model, or suboptimal feature representation.
- **Axes and Visual Interpretation**:

The x-axis represents the False Positive Rate (the proportion of incorrect positive predictions), and the y-axis represents the True Positive Rate (the proportion of correct positive predictions). The gray diagonal line (dotted) represents the performance level of a random model. Curves that approach this line indicate that the model does not add significant value over random prediction.

In conclusion, the classification model does not appear to perform well in predicting student participation in different sports disciplines.

- **AUC Values and Model Performance**:
  - The AUC (Area Under the Curve) values vary significantly across classes, ranging from 0.32 to 0.80.
  - **Class 11** shows the highest AUC (0.80), indicating that the model performs relatively well for this class, distinguishing it more effectively from the others.
  - **Classes with low AUCs**, such as Class 9 (0.32) and Class 0 (0.35), indicate that the model struggles to differentiate these classes from others, performing only slightly better than random guessing.
- **Inter-Class Variability:**
  - The variation in AUC scores suggests that the model's ability to predict participation differs significantly between classes. Some classes are easier for the model to identify (e.g., Class 11), while others are challenging.
  - This variability could imply that certain classes may have more distinct or discriminative features in the dataset, while others overlap more with other classes, making classification harder.
- **Curve Shapes and Classification Quality:**
  - The closer the curve is to the top-left corner, the better the model's performance for that class. In this image, only a few curves approach this corner, indicating suboptimal model performance overall.
  - Many curves closely follow the diagonal (dotted line), which represents random classification, indicating that the model performs poorly for those classes.
- **Potential Model Improvements**:
  - The model may benefit from further tuning, feature engineering, or a more advanced classification algorithm to improve its ability to distinguish between all classes.
  - Additional or more distinct data for classes with low AUCs might help the model learn to differentiate them better.

In conclusion, overall, the model shows moderate success in predicting participation for a few classes (like Class 11).

## CONCLUSION

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 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.

## REFERENCES

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. (2006). *Pattern Recognition and Machine Learning*. Springer.

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.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and self-determination of behavior. *Psychological Inquiry*, 11(4), 227-268.

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*(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.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260

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.

Scanlan, T. K., & Lewthwaite, R. (1986). Social psychological aspects of competition for male and female athletes. *Journal of Sport & Exercise Psychology*, 8(1), 53-71.

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

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

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 (Journal of Sport Psychology)*, *33*(2), 139-147.

Algerian University Sports Federation (n.d.). www.fasu-algeria.org., last accessed 20/11/2024.