ISSN 1870-4069

Predicting University Student Dropout with Extracurricular Activities Participation Using Machine Learning Models: A Case Study at Tecnológico de Monterrey

Francisco Mestizo, Alberto Orozco, Belén González,
Eunice Santos, Neil Hernandez-Gress

Tecnológico de Monterrey,
Monterrey, Mexico

{A01731549,A00831719,A01625378,A00831991,ngress}@tec.mx

Abstract. This study explores the impact of extracurricular activities on student retention at Tecnológico de Monterrey using data from the Institute for the Future of Education (IFE). The primary purpose of this study is to understand, predict, and prevent university dropout by examining the role of extracurricular activities and other relevant factors. By identifying the key determinants influencing dropout rates, we aim to offer actionable insights that can enhance student retention at a University. The research examines how participation in physical education, cultural activities, and student societies affects dropout rates among undergraduates. The initial analysis did not show a direct link between specific activities and retention, but higher overall engagement was correlated with reduced dropout rates. Machine learning models, including Support Vector Machines (SVM), Decision Trees, and Random Forests, were trained on a balanced dataset, with SVM achieving the highest accuracy at 61% after hyperparameter tuning. The study concludes that increased participation in extracurricular activities improves student retention, emphasizing the need for diverse programs to support student engagement. Also, the statistical analysis reveals that academic performance, financial support, and enrollment status are the primary predictors of student retention. These findings align with existing literature on student dropout, confirming the critical role of these factors. Future research should investigate additional non-academic factors and compare the findings between different institutions.

Keywords: Student dropout, machine learning, extracurricular activities.

1 Introduction

University dropout is a significant issue affecting educational institutions globally, including Mexico. High dropout rates impact students' future opportunities and the overall effectiveness of the educational system. Dropout can lead to financial instability, loss of human capital, and lower socio-economic mobility for individuals.

According to a report by the Organization for Economic Co-operation and Development [6], the dropout rates in higher education are a pressing concern, particularly in developing countries where education is a crucial pathway for social and economic advancement. Addressing this issue is vital for enhancing educational outcomes and ensuring that students can successfully complete their studies [5]. This study aims to explore how well student dropout can be explained by variables not extensively studied in previous research.

Specifically, we investigate the impact of extracurricular activities, along with other potential factors such as student engagement, socio-economic background, and academic support services. By examining these under-explored variables, we aim to provide new insights and contribute to the existing body of knowledge on student retention strategies. At Tecnológico de Monterrey, a variety of "LiFE" [1] courses and student groups are offered to enrich the student experience. These include sports teams, cultural clubs, leadership programs, and wellness activities.

These initiatives are designed to i) develop softskills and, ii) foster a sense of community, promote personal development, and enhance students' overall well-being. Understanding the role these activities play in student retention is crucial, as they are integral to the Tec's educational philosophy and mission to develop well-rounded individuals. The primary purpose of this study is to understand, predict, and prevent university dropout by examining the role of extracurricular activities and other pertinent factors. By identifying the key determinants influencing dropout rates, we aim to offer actionable insights that can enhance student retention at Tecnológico de Monterrey[2]:

– Research Question: Can the extracurricular activities offered by Tecnológico de Monterrey be a significant factor in preventing students from dropping out?

– Hypothesis: Extracurricular activities provided by Tecnológico de Monterrey play a relevant role in reducing student dropout rates.

This paper is organized as follows: Section 2 describes the State of the Art in Academic Performance and Student Dropout as well as Predictive Models. Section 3 describes the Database that has been developed [3] describing indicators and variables. Section 4 explains the Results of applying different Machine Learning methodologies to the Database. Finally Section 6 Concludes and discusses the future work.

2 State of the Art

2.1 University Dropout

University dropout has been extensively studied across various contexts, with numerous factors identified as key determinants. Among these factors, academic performance, socioeconomic status, extracurricular participation, and family support stand out as significant contributors.

[1] studyinmexico.tec.mx/es/life

[2] www.tec.mx

Academic Performance: Academic struggles, particularly in the first year, have been consistently linked to higher dropout rates. Students who fail to meet academic standards or experience difficulty adjusting to the academic rigors of university life are more likely to leave their studies prematurely. Rooij, Hansen, and Grift(2018) [12] found that first-year academic performance was a strong predictor of student retention in universities. Their research highlighted that early academic success is crucial for maintaining student engagement and persistence in higher education.

Socioeconomic Status: Socioeconomic challenges, including financial instability and the need to balance work with studies, are critical factors influencing dropout rates. Students from lower-income families often face additional pressures that can impact their ability to remain enrolled. Financial aid and scholarship programs have been found to play a crucial role in mitigating these challenges and improving retention rates. Aina, Baici, [1] demonstrated that socioeconomic status significantly affects student retention and graduation rates, emphasizing the need for robust financial support systems to aid disadvantaged students.

Extracurricular Participation: Engagement in extracurricular activities, such as sports, cultural clubs, and leadership programs, has been positively associated with student retention. These activities help students build a sense of community and belonging, which can be vital for their overall well-being and academic success. [8] highlighted the importance of extracurricular engagement in enhancing student retention and success. Their study showed that students who participate in extracurricular activities are more likely to stay enrolled and achieve academic success.

Family Support: The level of family support, including parents' educational background and involvement in their children's education, significantly affects student persistence. Research indicates that students whose parents have higher educational attainment levels are more likely to complete their studies. Werfhorst, [7] explored the impact of parental education on student persistence in higher education, finding that higher parental education levels correlate with increased student retention rates.

2.2 Predictive Models

Machine learning models have gained prominence in general and more specific in predicting student dropout due to their ability to handle large datasets and identifying complex patterns. Several models have been utilized to predict student retention, each with its strengths and limitations. Decision Trees: Decision trees are a non-parametric supervised learning method used for classification and regression. They provide a visual representation of decision rules and are easy to interpret.

However, decision trees can be prone to over-fitting, especially with small datasets. [2] applied decision tree analysis to identify at-risk students and found it effective in highlighting key predictors of dropout. Their study underscored the importance of decision trees in educational data mining due to their simplicity and interpretability. Random Forests: Random forests, an ensemble learning method, improve upon decision trees by constructing multiple trees during training and outputting the class that is the mode of the classes of individual trees.

This method reduces overfitting and increases accuracy, making it a robust choice for predictive modeling. [10] found that random forests provided superior performance in predicting student dropout compared to other machine learning methods. Their research highlighted the robustness and accuracy of random forests in educational settings. Support Vector Machines (SVM): SVMs are powerful for classification tasks, particularly when the data is high-dimensional. They work by finding the hyperplane that best separates the data into classes. SVMs are effective in capturing complex relationships between variables, though they require careful tuning of parameters.

In one study [9], the main predictors that are used to estimate student dropout are: Regime, Application grade, Internet, Car, Educational level of Parents, Occupation, Private insurance, Part-Time Job, Desktop Computer, Laptop. The results of this investigation lead to an accuracy approximate to 76.8%. [9] Another study (biblio) related with student dropout prediction using Machine Learning focuses on independent variables that are more related to student metrics about their own performance in school, mood during days of class and a few personal features: Gender, Former education level, Application priority, Degree program, Mother language, Start Semester, Age at enrollment, Credits, GPA, Failed Courses, exchange days, Moodle count, Moodle trent.

This study obtained results that were approximate to an accuracy of 76% determining the output if the student graduated or not[11]. While these researches focus on specific attributes to determine student dropout, like the student's performance in school, mood, socioeconomic and some other personal characteristics, this research opted for using the same approach but with a different focus. Most of the variables that were considered are related to the student's involvement in extracurricular activities, so we could determine dropout based on the influence of these activities.

3 Experimental Development

In order to have a better understanding of the proposed methodology for solving the designated challenge, it is necessary to establish some basic concepts. For this, it is essential to talk about Artificial Intelligence itself, more specifically, Machine Learning. This is described as the use of computational methods to be trained from data without the need of managing manual control in the applied procedure. It is to say that artificial intelligence models are being used to solve problems of almost any nature automatically. [4]

For this research, multiple options are considered in order to find the best model possible with the best metrics. In Machine Learning, exist multiple models that can provide good estimations for student dropout in college, like the Support Vector Machine, Neural Networks, Decision Tree and Random Forests to name a few related to supervised learning, also there are models for clustering like K-Means or DBSCAN used in unsupervised learning.

Based on the research performed, there were found interesting methodologies applied for very similar problems that were carried out with Machine Learning. In one study there was mentioned that supervised learning has the most accurate results with three main models that have proven to be the most outstanding ones referring to achieving a robust classification performance and distinguished accurate results [13].

The first one is the Support Vector Machines model, this algorithm is capable of performing classification tasks such as binary and multi-class classification, as well as regression problems. The objective of this model is finding the hyperplane that better divides the data classes. Some of the advantages of applying decision trees are:

– Decision trees are easy to understand, interpret and visualize.

– Can handle categorical as well as numeric variables without the need of complex pre-processing.

– It doesn't require data normalization or feature standardization.

– Captures non.linear relations between features and tags.

4 Results

The data that supports the findings of this study are available from the Institute for the Future of Education (IFE)'s Educational Innovation collection of the Tecnológico de Monterrey's Research Data Hub but restrictions apply to the availability of these data, which were used under a signed Terms of Use document for the current study, and so are not publicly available. However, data is available from the IFE Data Hub upon reasonable request at https://doi.org/10.57687/FK2/PWJRSJ (accessed on 11 April 2024). The used database has information from Tec de Monterrey's high school and undergraduate students [3]. The dataset has information related to academic performance, participation in extracurricular activities inside the school and information from the parents of the students. To reduce the scope for this research, it is only focused in the undergraduate students. The description of each category will be given below:

1. Physical Education: Students enrolled in one or more physical activities. One in this column represent that the student is enrolled and zero means they are not enrolled.
2. Cultural Diffusion: Students enrolled in one or more cultural activities. One in this column represent that the student is enrolled and zero means they are not enrolled.
3. Student Society: Students actively involved in one or more student groups. One in this column represent that the student is involved and zero means they are not involved.
4. Total LiFE Activities: LiFE activities are the extracurricular activities offered by Tec de Monterrey. This column indicates in how many activities the student is enrolled.
5. Athletic Sports: Students that are part of any team that represent the university in different sports competition.
6. Art Culture: Students that are part of any team that represent the university in different cultural activities, such as theater and dance.
7. Student Society Leadership: Students that are part of the board of any student society.
8. LiFE Work Mentoring: Student that are part of programs like Peer Mentor, where they help freshmen and sophomores.

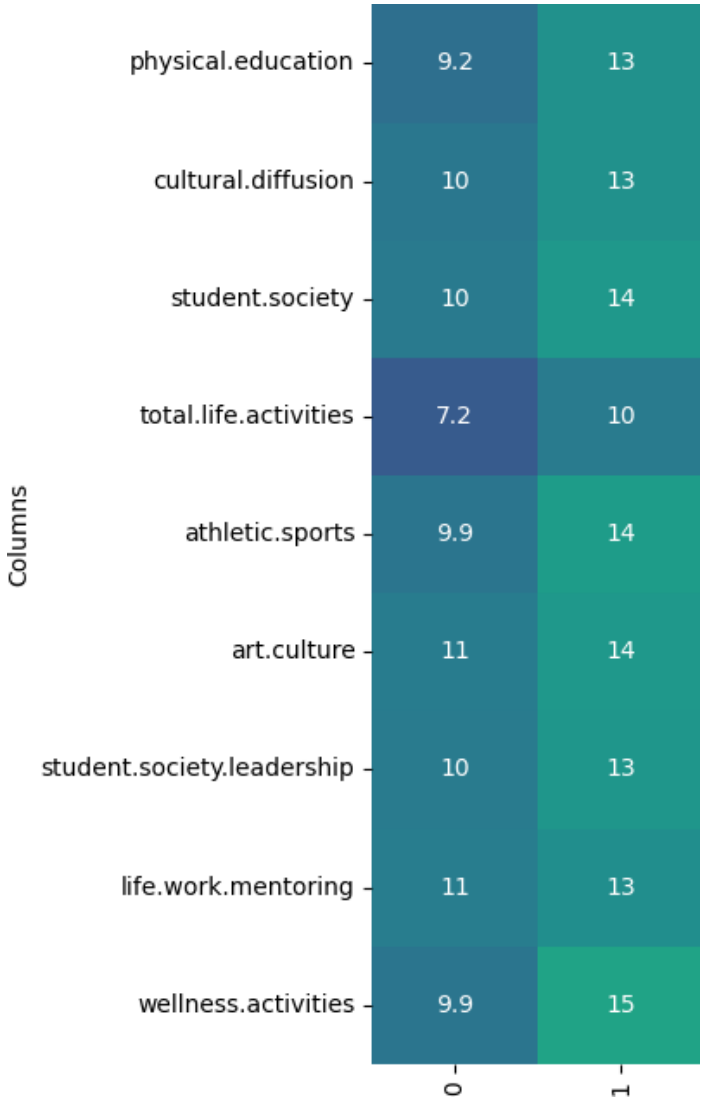

Fig. 1. Ratio of school permanence by activity.

9. Wellness Activities: Students enrolled in well-being activities, like meditation.
10. Foreign: Indicator to identify if the student is a foreigner (Yes: Foreigner), if the student's birthplace is different from the location of the school campus (Yes: National) or belongs to the same location (Local).
11. PNA: Previous level score (Average).

The value to be predicted is represented in the database as Retention. A value of one means the student is still studying, while a value of zero means they left school.

4.1 Data Cleaning

To start using the data it has to be cleaned. Overall, it is a well-structured database, as it does not have any null value for any column. Also, most of the columns are self-described so it is easy to understand what they are measuring just by reading the name. Some of the columns contains data as "No information" or "Does not apply". For this data to be used for the models, that values where changed for 0. For the categorical columns as foreign, one hot encoding was performed.

4.2 Data Analysis

After the columns where selected, a correlation analysis with retention was conducted to understand the impact these variables have in retention. All the variables had a correlation coefficient of less than 5%, so the variables do not have direct correlation with retention. It was analysed how much impact each category had by itself. This was done by comparing all the students that where on a physical activity and dropped with all the students that were not in a physical activity compared to the students that were on a physical activity and stayed in school. With that result, a ratio was calculated.

This was done with each variable. The result showed a biggest ratio of participation in any activity for students that did not quit school. The difference in these ratios is on average of three points as shown in figure 1. For total life activities, a curious pattern is shown, where having one extra LiFE activity increments by average three points the ratio. This means that the more LiFE activities an student has, the least probable it is that they quit school. So, even though this categories did not show direct correlation with retention by themselves, their presence appears to be important.

4.3 Data Sampling

The dataset contains information from all the students from generation 2014 through 2020. Just for this generations, there are 77,517 students in the database. From those, 70,704 stayed in school while the other 6,813 quit school. It is reassuring to see that the number of dropouts is much lower than the students that stayed, but it represents a problem for training a machine learning model [13]. The desired dataset for training a machine learning model is to have balanced classes, this means to have similar amount of data for each class. There are several methods to accomplish this, like applying data augmentation by generating random data from dropouts to balance both classes. Another approach, which is the one taken on this research, is to take all dropout registers and take randomly the same amount for students that stayed on school. This generates a final dataset with 50% - 50% distribution of the classes.

4.4 Model Training and Selection

According to Villar [13], the models that have shown best results for predicting student dropout are Support Vector Machines (SVC), Decision Trees and Random Forest. No hyperparameters will be used for this stage. The models will be trained using the default values for their hyperparameters.

Those three models were trained with the same balanced dataset to compare the performance of each one based on accuracy. Then, the best one will be selected for a refinement of its hyperparameters. SVM Model: 0.58, Decision Tree: .54 and Random Foret 0.54 The results for each model can be seen in the tables 1 through 3. On average, the results of accuracy for the models are 56%, with the best model being SVM with 59% of accuracy. The results are not optimal, but they may be improved by refining the model.

4.5 Model Refinement

Some of the hyperparameters that can be defined for a SVM model are:

- C (Regularization parameter): Controls the trade-off between achieving a low error on the training data and minimizing the complexity of the model (which helps in avoiding over-fitting).
- Gamma: Defines how far the influence of a single training example reaches.
- Kernel: Specifies the type of kernel to be used in the SVM algorithm. Linear means a linear boundary is used to separate the data, while Radial Basis Function (RBF) uses non-linear decision boundaries.

For the optimization, different combinations of those hyperparameters where tested. For C, the values where 0.1, 1 and 10; for Gamma, 0.1 1 and 10; and for kernel, linear and RBF. The values selected where arbitrary, just to test different states of the model. After being trained with the same dataset, the best model had the values. The results for accuracy are 0.58. The result do not show great improvement for the model. This could mean that this values cannot predict more than 60% of the data. As this is not a great result, it is quite acceptable for variables that were not that correlated to the predicted value.

4.6 Model Improvement

To test how much the model could be improved, the same steps where taken but using more categories. To select the new classes a Principal component analysis (PCA) and correlation analysis was done.

Principal Components Analysis. Principal Component Analysis (PCA) was performed to reduce the dimensionality of the data and identify the main components contributing to variance. The first two principal components explain approximately 32.2% of the total variance, indicating that while these components capture a significant portion of the data's variability, other factors also contribute to the overall dropout rates. These findings indicate that academic performance (Average First Period Grade) is the most significant predictor of student retention. This aligns with the general understanding that students with better academic performance are less likely to drop out. Financial support (Scholarship Percentage and Total Scholarship Loan) and enrollment status (FTE) also play crucial roles, suggesting that students who receive financial aid and are fully enrolled are more likely to persist in their studies.

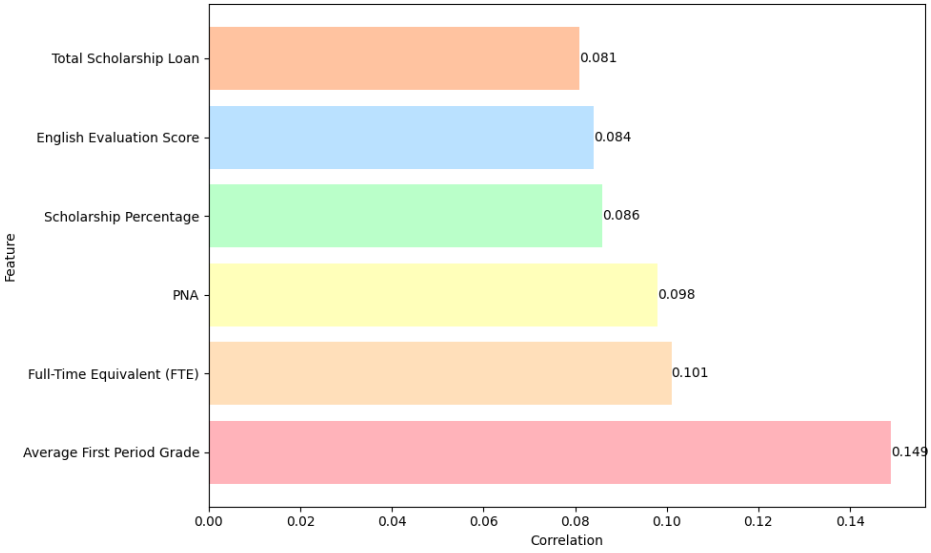

Fig. 2. Top correlated features with retention.

The PNA and English Evaluation Score reflect engagement and proficiency, further underlining the importance of academic and support structures.

Correlation Analysis. The correlation between each feature and the target variable (retention) was computed. The top six correlated features are shown in Figure 2. The correlation analysis supports the findings from the feature importance analysis. High correlations between retention and features like Average First Period Grade and Scholarship Percentage emphasize the critical role of academic performance and financial support in preventing dropout.

4.7 Model Re-training

After the mentioned analysis, the added categories were:

1. English Evaluation: Level of English obtained from a standardized English proficiency test.
2. Tec no Tec: Indicator that denotes if the student comes from a school that belongs to Tecnológico de Monterrey.
3. Parents Exatec: Indicator that denotes if either one of the two parents is exatec (was a student of Tecnológico de Monterrey).
4. General Math Eval: Mathematics grade from the admission test or from the school of origin.

Three of this categories are related to academic performance, while the other is centered in the school background of the family.

For this training, the same data sampling and cleaning methods were used. The only difference was for the General Math Eval column. For the values of "No information" and "Does not apply" zeros were not added as for the other columns. That could cause problems for the predictions because 0 have a different meaning in this column. That's why the mean of the other values of the column was obtained and then changed for the registers that did not had a numerical value. With the addition of these the variables we can see a slightly increase in accuracy, but it doesn't seem to be too significant. The best model was SVM again, with an accuracy of 61%. The result for other models are: Decision Trees: 0.55, Random Forest: 0.56.

5 Analysis

The statistical analysis reveals that academic performance, financial support, and enrollment status are the primary predictors of student retention. These findings align with existing literature on student dropout, confirming the critical role of these factors. But, extracurricular activities, although not among the top predictors, show a noticeable impact on retention, as shown in this research. This supports the hypothesis that these activities foster engagement and a sense of belonging, which are essential for student persistence. The findings are noticeable by the dataset from Tecnológico de Monterrey, which includes detailed information on academic performance and financial support but also from extracurricular involvement.

The use of advanced machine learning models, particularly Support Vector Machines, allowed uncovering these insights with high accuracy. The study demonstrates the effectiveness of using machine learning models to predict student dropout and identify key factors influencing retention, even thought the variables had no correlation at all. The findings suggest that institutions should focus on enhancing academic support, providing financial aid, and promoting extracurricular activities to improve student retention. Future research could explore the specific types of extracurricular activities that have the most significant impact and develop tailored interventions to support at-risk students.

5.1 Student Retention through Extracurricular Programs

Based on the results, its proposed that other institutions adopt and expand extracurricular activities similar to those offered at Tecnológico de Monterrey. Programs like LIFE, which integrate leadership, innovation, and entrepreneurial training with traditional extracurricular activities, could be particularly beneficial. These programs not only enhance student engagement but also build skills that contribute to academic and professional success. Below are examples of extracurricular activities that institutions could adopt:

– Leadership Programs: Workshops and seminars that develop leadership skills and provide opportunities for students to take on leadership roles.
– Innovation and Entrepreneurship: Incubators and hackathons that encourage creative problem-solving and entrepreneurial thinking.

– Sports and Fitness: Comprehensive sports programs that promote physical health and team-building skills.
– Cultural and Arts Programs: Activities that celebrate cultural diversity and foster creative expression.
– Community Service: Volunteer opportunities that help students develop empathy and a sense of social responsibility.
– Professional Development: Career counseling, internships, and networking events that prepare students for post-graduation success.

6 Conclusion and Future Work

Support Vector Machines, as seen in other bibliography, resulted in one of the most effective machine learning models to predict student dropout. The use of variables such as students' extracurricular performance, although not strongly correlated with their retention, proved to be variables that could predict up to 61% of the cases of students who remained in or left school. From this research, it is encouraged to use machine learning models in other variables not related to academic performance, such as extracurricular performance, geographical context, health, family, technological development or emotional state. Also, it could be interesting to do this same analysis in other schools near Tecnológico de Monterrey to see if the obtained results are similar or do not.

Acknowledgments. The authors would like to thank Tecnológico de Monterrey and the Living Lab and Data Hub of the Institute for the Future of Education for the data provided for this research.

References

1. Aina, C., Baici, E., Casalone, G., Pastore, F.: The determinants of university dropout: A review of the socio-economic literature. Socio-Economic Planning Sciences, vol. 79, pp. 101102 (2022) doi: 10.1016/j.seps.2021.101102
2. Albreiki, B., Habuza, T., Zaki, N.: Extracting topological features to identify at-risk students using machine learning and graph convolutional network models. International Journal of Educational Technology in Higher Education, vol. 20, no. 1 (2023) doi: 10.1186/s41239-023-00389-3
3. Alvarado-Uribe, J., Mejía-Almada, P., Masetto-Herrera, A. L., Molontay, R., Hilliger, I., Hegde, V., Montemayor-Gallegos, J. E., Ramírez-Díaz, R. A., Ceballos, H. G.: Student dataset from Tecnológico de Monterrey in Mexico to predict dropout in higher education. Data, vol. 7, no. 9, pp. 119 (2022) doi: 10.3390/data7090119
4. Bhasin, H.: Machine learning for beginners: Learn to build machine learning systems using python. BPB Publications (2020)
5. Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., Paganoni, A. M.: Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. Studies in Higher Education, vol. 47, no. 9, pp. 1935–1956 (2021) doi: 10.1080/03075079.2021.2018415

6. Canton, H.: The Europa directory of international organizations. Routledge (2021) doi: 10.4324/9781003179900
7. Forster, A. G., van-de-Werfhorst, H. G.: Navigating institutions: Parents' knowledge of the educational system and students' success in education. European Sociological Review, vol. 36, no. 1, pp. 48–64 (2019) doi: 10.1093/esr/jcz049
8. King, A. E., McQuarrie, F. A., Brigham, S. M.: Exploring the relationship between student success and participation in extracurricular activities. SCHOLE: A Journal of Leisure Studies and Recreation Education, vol. 36, no. 1–2, pp. 42–58 (2020) doi: 10.1080/1937156x.2020.1760751
9. Sandoval-Palis, I., Naranjo, D., Vidal, J., Gilar-Corbi, R.: Early dropout prediction model: A case study of university leveling course students. Sustainability, vol. 12, no. 22, pp. 9314 (2020) doi: 10.3390/su12229314
10. Utari, M., Warsito, B., Kusumaningrum, R.: Implementation of data mining for drop-out prediction using random forest method. In: Proceedings of the 8th International Conference on Information and Communication Technology, pp. 1–5 (2020) doi: 10.1109/ICoICT49345.2020.9166276
11. Vaarma, M., Li, H.: Predicting student dropouts with machine learning: An empirical study in finnish higher education. Technology in Society, vol. 76, pp. 102474 (2024) doi: 10.1016/j.techsoc.2024.102474
12. van-Rooij, E. C. M., Jansen, E. P. W. A., van-de-Grift, W. J. C. M.: First-year university students' academic success: The importance of academic adjustment. European Journal of Psychology of Education, vol. 33, pp. 749–767 (2017) doi: 10.1007/s10212-017-0347-8
13. Villar, A., de-Andrade, C. R. V.: Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. Discover Artificial Intelligence, vol. 4, no. 1 (2024) doi: 10.1007/s44163-023-00079-z