

Algorithmic Prediction of Students On-Time Graduation from the University

Ayman Alfahid¹

¹ *Majmaah University, Industrial Area, Riyadh, Saudi Arabia*

Abstract – This study develops statistical learning models to assess the probability of undergraduate students graduating within a predetermined period, utilizing admission, performance, and demographic data. The urgency of addressing student attrition is highlighted by recent data from the National Center for Education Statistics (NCES), indicating a 59% completion rate by full-time undergraduates within six years. This research leverages institutional data from a Saudi University, focusing on freshmen enrolled in the 2012-2013 and 2013-2014 academic years, to identify students at risk of dropping out, thereby enabling timely interventions. Ten algorithms, including decision trees, ensemble models, SVM, and ANN, were built and evaluated on a test set representing 33.3% of the entire dataset using precision, recall, accuracy, and Matthews correlation coefficient (MCC). The findings show that SVM and Random Forest models were the most reliable, achieving accuracies of 0.830 and 0.831 respectively, and maintaining balance in precision, recall, and MCC. Conversely, the naïve Bayes model recorded the worst performance. The comparative analysis revealed the superior performance of ensemble models over decision tree models in predicting student attrition, emphasizing the importance of model selection in developing effective early intervention strategies. In addition, our analysis revealed that academic data is a better predictor of on-time graduation than admission data, emphasizing the need for institutions to focus on continuous academic assessment data.

Keywords – Student dropout, ensemble models, random forest.

1. Introduction

Student dropout is a critical concern in higher education, affecting institutional reputations, financial stability, and student success. The National Center for Education Statistics (2017) reported that about half of those who enroll in higher education complete their Bachelor's degree. Porter [17] noted that 40% of college students leave higher education without earning a degree. This high drop-out rate has necessitated a focus on student retention strategies, especially in institutions in the United States [10], where it has become a priority for administrators.

The first year of college is pivotal; research shows a substantial correlation between first-year academic performance and retention rates [5]. Many students tend to withdraw during this initial phase of their academic journey. Consequently, early identification of students at risk of dropping out is crucial. It enables institutions to implement targeted intervention strategies, providing at-risk students with appropriate resources and guidance to improve their chances of success [6], [22]. Reports from early warning systems can be shared with administrators, teachers, and academic supervisors to help identify and support students at risk [14]. Previous research [12] has looked into the generalizability and accuracy of predictive models in the context of distance learning systems and emphasized the need for further research in different learning environments.

Therefore, this study aims to advance the state of the art towards guiding decision-makers in educational institutions by providing insights into student retention, utilizing institutional data to pave the way for effective early intervention strategies and, subsequently, enhancing student success rates. The research questions are as follows:

1. How effective are different classification algorithms in predicting student attrition at King Saud University?
2. Do ensemble models offer superior performance over decision tree models in predicting student attrition for early interventions?

DOI: 10.18421/TEM131-72

<https://doi.org/10.18421/TM131-72>


Corresponding author: Ayman Alfahid,
Majmaah University, Industrial Area, Riyadh, Saudi Arabia
Email: e.alfahed@mu.edu.sa

Received: 26 September 2023.

Revised: 23 October 2023.

Accepted: 31 January 2024.

Published: 27 February 2024.

 © 2024 Ayman Alfahid; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

2. Literature Review

Machine learning techniques have established themselves as pivotal tools in the administration and management of post-secondary education institutions. The supremacy of their application in enrollment management and student retention has been corroborated by various studies, highlighting the efficacy of these techniques in providing substantial insights for policymakers [14], [15], [4]. The studies, although varying in objectives and considered variables, uniformly validate the precision of machine learning techniques in predicting student performance.

2.1. Predictors of Student Dropout

Post-secondary institutions in the United States have seen a steady increase in enrollments. However, problems like poor academic performance and high dropout rates among undergraduates continue to persist [1]. Many researchers have focused on finding out which students are at risk and what help they need, while some have focused on the factors that keep students enrolled [1].

Different studies have explored the myriad factors influencing dropout rates. Paterson [16] in his study, identified scholarship status and matriculation as main factors of retention at the West Indies University. He referred to theories like Astin's theory of student involvement, which relates students' time and involvement in activities to their engagement level in school activities.

A study involving various Spanish universities [2] identified start age, academic performance, parental education level, and number of attempts required to pass as significant determinants of university dropouts. The distinct profiles of students likely to drop out are found to be contingent on the subject or course studied.

Also, Delen [6] noted that despite extensive student retention efforts, about half of the entrants in higher education in the United States abandon their programs before attaining a Bachelor's degree. He contended that retention strategies necessitate a profound understanding of the underlying causes of dropout.

Regarding distance education, the absence of in-person teacher interaction can lead to a sense of isolation among students, causing distress and subsequently, withdrawal from courses [12]. To mitigate these concerns, multiple studies have advocated the incorporation of machine learning techniques.

These techniques are aimed at predicting potential dropouts at various stages of a course and pinpointing the signs indicative of dropping out [16].

2.2. The Role of Learning Analytics

Cohen [3] proposed that by monitoring student activities in web-supported courses, educators could potentially foresee student dropouts. This approach is pivotal for identifying students at risk of dropping out in a timely manner. Several studies have validated that learning analytics, the evaluation of students' online data, play a critical role in refining learning processes and maximizing the efficiency of educational environments [22]. Such analytical insights can serve as a proactive warning system, allowing for interventions to be implemented promptly.

The Learning Management System (LMS) has also been acknowledged as a useful tool for identifying students who are at risk, facilitating the execution of prompt interventions. In their research, McFadyen and Dawson [14] analyzed students' online activities that significantly correlate with academic achievement. Through regression modeling, they developed an optimized predictive model to elucidate the variations observed in students' final grades. By analyzing course discussion forums, they demonstrated how such forums offer insights into the development of student learning communities by recognizing isolated students, analyzing student-to-student interaction patterns, and determining the role of the instructor within the network.

2.3. Predictive Models

Various research, including the work of Marquez-Vera *et al.* [16], have conducted comparative analysis on different classification methods such as decision trees, neural networks, support-vector machines, and logistic regression. These models were pivotal in developing models to predict at-risk first-year students. The challenges encountered during the classification of at-risk students were mitigated by implementing data resampling using SMOTE and cost-sensitive learning [16].

Herzog [9] contends that machine learning techniques yield higher prediction accuracies relative to other methods. This assertion is rooted in the study's comparison of the predictive accuracy of three artificial neural networks and three decision trees with that of multinomial logistic regression.

3. Methods

This research is a binary classification task with the target being “Attrition Status”, categorized as “No” if a student graduates within four years and “Yes” otherwise. The R programming language, version 4.2.2 [19], is used for experiments.

3.1. Dataset

The study focuses on undergraduate students at a Saudi University. The dataset is sourced from the University’s Student Information System (SIS), with preliminary access approval granted by the University. A subset of students enrolled during the academic years of 2012-2013 and 2013-2014 forms the sample for this study. The features are divided into four categories: student demographic data (i.e., birth year, age, and gender), students’ admission data (i.e., high school GPA, graduation year, and district), students’ aptitude data, and students’ academic data (i.e., cumulative GPA, last semester GPA, and total earned hours).

Figure 1 shows the percentage of students’ on-time graduation per college.

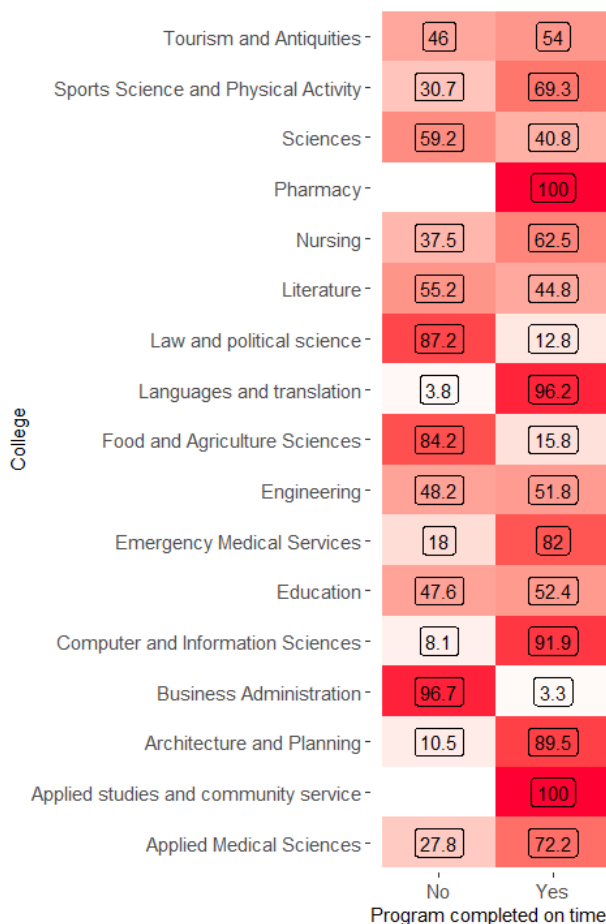


Figure 1. Percentage of Students’ on-time Graduation per College

3.1.1. Data Preprocessing

During the data preprocessing phase, we dropped rows with missing values. Also, the 'Gender' feature and the target variable 'Graduate on time?' were binary encoded. Although the 'Department' feature has 65 categorical values, we adopted one-hot encoding for it rather than numerical encoding to avoid any artificial ordinality that may be introduced via numerical encoding. Also, the ‘Department’ feature was selected over college because of the granularity it offers in terms of academic disciplines.

The class distribution of the target variable 'Graduate on time?' showed a balanced distribution with 51.24% labeled 'False' and 48.76% 'True'. Given this nearly even split, there was no need for further class imbalance mitigation. The dataset was pre-processed by dropping instances with missing values, feature engineering, etc. Ultimately, the dataset has 5,883 instances and 10 features presented in Table 1.

Table 1. Features selected for classification

Category	Features	Range
Demographic	Gender	Male or Female
Admission	General aptitude test	0 – 100
	High school branch	5 branches
	Achievement test	0 – 100
	High school GPA	0 – 100
Academic	Department	65 Depts
	Plan hours	128-185 hrs
	Hours registered in last semester	0 – 25 hours
	First Year Average	0 – 5
	Program Length	8 – 14 semesters

3.2. Model Selection

In this research, a total of 10 classification algorithms are implemented to assess completion time, with the detailed list and their corresponding types provided in Table 2 below:

Table 2. Selected Models

Model Type	Algorithms
Decision Tree	Classification and Regression Tree, Conditional Inference Tree
Ensemble	Bagging - Bagged Trees and Random Forest Boosting - XGBoost and AdaBoost
Others	ANN (Neural Network), Naïve Bayes (Probabilistic), SVM (Support Vector), and Logistic Regression.

1. Classification and Regression Tree (CART) and Conditional Inference Tree (CIT) are types of decision tree models that are valuable for their simplicity and ability to handle both categorical and numerical data.
2. Bagged Trees and Random Forest are ensemble models using bagging techniques. They operate by building multiple decision trees and merging them together to obtain a more stable and accurate prediction.
3. AdaBoost and XGBoost use boosting techniques to convert weak learners into strong ones. They are versatile and robust, offering solutions for both classification and regression problems.
4. Logistic Regression: Given a vector of input features x , the probability $P(Y = 1|X)$ that the target variable Y is 1 can be represented by the logistic function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

Where $\beta_0, \beta_1, \dots, \beta_k$ are the parameters of the model, and x_1, x_2, \dots, x_k are the input features.

5. Naïve Bayes classifier: Given a feature vector $x = (x_1, x_2, \dots, x_k)$, the posterior probability for a class C can be computed using Bayes' theorem.
6. Artificial neural networks (ANNs) are a category of models renowned for their capability to capture and represent intricate and non-linear relationships within data. These networks employ interconnected nodes or "neurons" organized in layers to learn from the input data progressively, adjusting the connections based on the error of the model's predictions.
7. Support vector machine: The decision function of a SVM for binary classification can be represented as:

$$f(x) = \beta_0 + \sum_{i=1}^n \alpha_i y_i x_i$$

Where α_i are the Lagrange multipliers, y_i are the class labels, x_i are the support vectors.

3.3. Model Evaluation

The performance of the models was evaluated on a test set, which is 33.3% of the entire dataset, using precision, recall, accuracy, and Matthews correlation coefficient (MCC). The MCC is particularly crucial as it offers insights into the correlation between predictions and actuals, with values closer to +1 indicating a more powerful model.

4. Results and Discussion

The results obtained for each model are presented in Table 3. From the results, it can be seen that SVM (Radial Kernel) and Random Forest models demonstrated the highest reliability, with accuracies of 0.830 and 0.831 respectively. These models maintained a balance in precision, recall, and MCC, indicating consistent performance in different aspects of classification. The effectiveness of the random forest model can be attributed to its capability to manage non-linearities and interactions between features, making it suitable for the complexities of our dataset. Similarly, the SVM model, with its radial kernel, managed the non-linear decision boundary effectively, contributing to its high performance.

In contrast, the naïve Bayes model was the least reliable, with an accuracy of 0.703 and a precision of 0.655. Despite its high recall of 0.894, the model's tendency to produce a higher number of false positives questions its accuracy in predicting student at risk of dropping. This suboptimal performance may be due to its assumption of feature independence, which may not align with the realities of educational data, where features can be interdependent.

4.1. Ensemble Models vs Decision Trees

Comparing ensemble models and decision trees, it is clear that ensemble models performed better on average than the single decision tree model, CART. Ensemble models, including Random Forest, Bagged CART, XGBoost, and AdaBoost, showed balanced accuracy, precision, and recall, with Random Forest as the top performer among ensemble models. The robustness of ensemble models is due to their ability to aggregate predictions from multiple models, reducing the risk of overfitting found in single decision tree models like CART.

4.2. Bagging vs Boosting

In the comparison of ensemble models, bagging models, represented by Random Forest and Bagged CART, show a slight edge over boosting models like XGBoost and AdaBoost. Bagging models are valued for their capability to reduce overfitting and variance, which is crucial for handling unstable models and enhancing their performance in classification tasks. Boosting models, while also effective, focus on minimizing bias and improving accuracy by combining the outputs of multiple weaker models.

This slight advantage of bagging models underscores the importance of meticulously aligning model selection with the dataset's characteristics and the goals of the research.

Table 3. Classification results obtained for all ten algorithms

MODEL	OPTIMAL PARAMETERS	TUNE LENGTH	ACCURACY	PRECISION	RECALL	MCC
CART	CP = 0.0140	25	0.797	0.832	0.758	0.596
CIT	MINCRITERION = 0.4183	25	0.819	0.877	0.754	0.645
SVM (RADIAL KERNEL)	SIGMA = 0.0360 C = 4	12	0.830	0.877	0.778	0.665
GLM	(N/A)	(N/A)	0.823	0.854	0.792	0.648
ARTIFICIAL NN	SIZE = 5, DECAY = 0.0251	12	0.814	0.912	0.706	0.647
NAÏVE BAYES	USEKERNEL = TRUE	2	0.703	0.655	0.894	0.431
BAGGED CART	(N/A)	(N/A)	0.810	0.831	0.792	0.621
RANDOM FOREST	MTRY = 6, NTREE = 128 (FIXED)	12	0.831	0.870	0.789	0.665
ADABOOST	NITER = 50, ADABOOST.M1	6	0.809	0.832	0.789	0.619
XGBOOST	NROUNDS = 50, MAX_DEPTH = 4, ETA = 0.3, GAMMA = 0, COLSAMPLE_BYTREE = 0.6, MIN_CHILD_WEIGHT = 1, SUBSAMPLE = 0.7	6	0.823	0.851	0.795	0.647

4.3. Model Interpretability

Model interpretability offers a clear understanding of feature importance in predictions, thereby increasing trust in the model's outcomes. To this end, we employed both SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to

understand the influence of features in our best-performing model, Random Forest.

Global Explanation with SHAP: A global perspective was achieved using the SHAP plot. In the context of our model, the SHAP plot (Figure 2) uncovers the overall significance of each feature across all predictions.

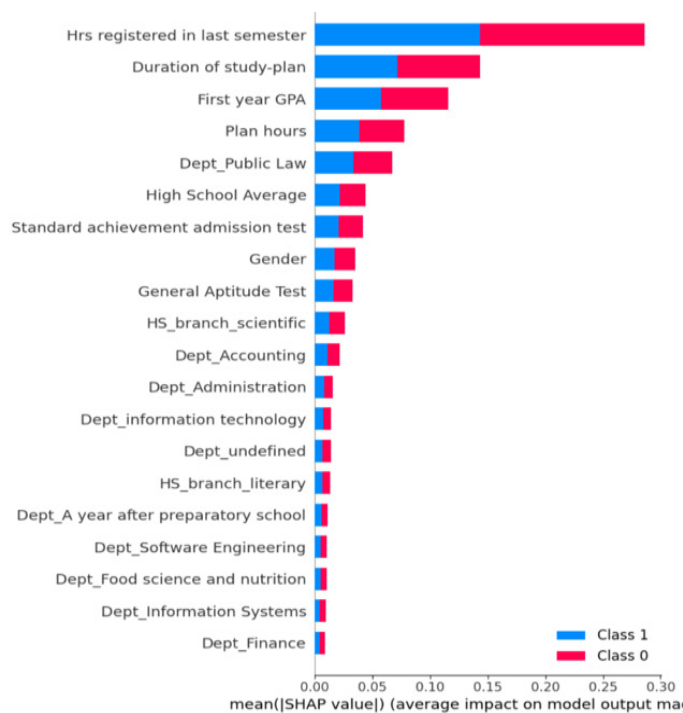


Figure 2. Global Explanation of predictions using Shapley

It indicates that the most influential features are 'hrs registered in last semester', 'duration of study plan', 'first year GPA', 'plan hours', 'dept public law', 'high school average', and 'standard achievement admission test'. These features consistently had the highest impact on the model's decision-making process, emphasizing their role in determining whether a student will graduate on time or not.

Local Explanation with LIME: For a more granular, instance-specific understanding, LIME was implemented on data instance 900 (Figure 3). LIME facilitates the interpretation of individual predictions by approximating the model's behavior in the neighborhood of the chosen instance. The analysis for instance 900 highlighted 'duration of study plan' and 'hours registered in the last semester' as the most influential features. This particularly aligns with the insight gotten from the global explanation.

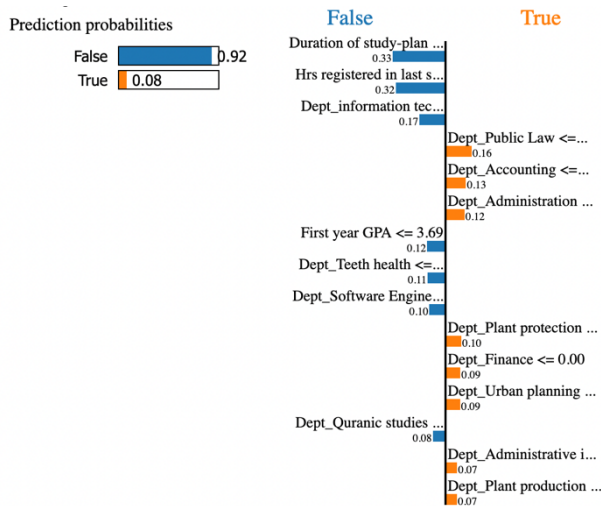


Figure 3. Local explanation of prediction for Instance900

4.4. Admission vs Academic Data

In the analysis of the predictive capabilities of academic versus admission data on the likelihood of a student graduating on time, distinct variations in performance metrics were observed. Using our best performing model, RandomForest, the academic dataset yielded a substantially higher accuracy, precision, recall, and f1-score, all around 80%, while the admission dataset performed poorly with an accuracy of only 56% and similar metrics hovering in the mid-50s range. The results are presented in Tables 4 and 5.

Table 4. Results obtained for academic data

	precision	recall	f1-score	support
0	0.80	0.81	0.80	894
1	0.80	0.79	0.79	871
accuracy			0.80	1765
macro avg	0.80	0.80	0.80	1765
weighted avg	0.80	0.80	0.80	1765

Table 5. Results obtained for admission data

	precision	recall	f1-score	support
0	0.56	0.58	0.57	894
1	0.55	0.54	0.55	871
accuracy			0.56	1765
macro avg	0.56	0.56	0.56	1765
weighted avg	0.56	0.56	0.56	1765

The comparatively superior performance of the academic data emphasizes the argument that educational institutions might benefit from waiting until students complete their first year to achieve a more accurate prediction of on-time graduation. The first-year GPA, for instance, can serve as a crucial indicator of a student's adaptability, commitment, and potential trajectory through their academic journey. Conversely, the limited predictive capability of the admission data suggests that while these metrics are useful for initial student selection, they may not be the most optimal for long-term academic outcome predictions.

In summary, for institutions aiming to harness data-driven strategies to optimize student success, the focus might need to shift towards leveraging continuous academic assessments.

5. Conclusion

This study conducted a thorough evaluation of several machine learning models to predict student completion rates within a set timeframe, using metrics such as accuracy, precision, recall, and MCC for a well-rounded analysis. The SVM (Radial Kernel) and random forest models proved to be the most reliable, displaying balanced performance across all metrics. Conversely, the naïve Bayes model showed lower reliability, marked by a higher number of false positives. Additionally, the research underscored the effectiveness of ensemble models, especially those like random forest that use bagging techniques, in balancing interpretability and predictive accuracy, outperforming single decision tree models and offering valuable insights for shaping educational strategies. Another revelation in this study was the comparative predictive strength of academic data over admission data.

Meanwhile, a critical aspect that warrants attention in future studies is the consideration of the fairness of algorithms, especially in relation to the source of the data used. The exploration of the accuracy-fairness tradeoff is essential to understand the implications of using different types of data, such as admission data versus academic data. Evaluating whether it is fairer to use one type of data over the other and understanding the ethical and practical ramifications of these choices are paramount. This exploration can contribute to the development of models that are not only accurate and reliable but also equitable and ethical, fostering a more inclusive and fair educational environment.

References:

- [1]. Miller, T. E., & Herreid, C. H. (2009). Analysis of variables: Predicting sophomore persistence using logistic regression analysis at the University of South Florida. *College and University*, 85(1), 2.
- [2]. Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university dropout rates. *Computers & Education*, 53(3), 563–574. Doi: 10.1016/j.compedu.2009.03.013
- [3]. Cohen, A. (2017). Analysis of student activity in web-supported courses as a tool for predicting dropout. *Educational Technology Research and Development*, 65(5), 1285–1304. Doi: 10.1007/s11423-017-9524-3
- [4]. Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), 17–29. Doi: 10.1109/tlt.2016.2616312
- [5]. DeBerard, M. S., Spielmans, G., & Julka, D. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal*, 38(1), 66–80.
- [6]. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498–506. Doi: 10.1016/j.dss.2010.06.003
- [7]. Freund, Y., & Schapire, R. E. (1995, March). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* 23-37. Berlin, Heidelberg: Springer Berlin Heidelberg.
- [8]. Gansemer-Topf, A. M., & Schuh, J. H. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in higher education*, 47, 613-642.
- [9]. Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 2006(131), 17–33. Doi: 10.1002/ir.185
- [10]. Ishitani, T. T. (2016). Time-Varying Effects of Academic and Social Integration on Student Persistence for First and Second Years in College. *Journal of College Student Retention: Research, Theory & Practice*, 18(3), 263–286. Doi: 10.1177/1521025115622781
- [11]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer. Doi: 10.1007/978-1-4614-7138-7
- [12]. Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, 23(6), 529–535. Doi: 10.1016/j.knsys.2010.03.010
- [13]. Kuhn, M. (2022). *caret: Classification and Regression Training. R package version 6.0-93*. CRAN.R. Retrieved from: <https://CRAN.R-project.org/package=caret> [accessed: 29 August 2023].
- [14]. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an 'early warning system' for educators: A proof of concept. *Computers & Education*, 54(2), 588–599. Doi: 10.1016/j.compedu.2009.09.008
- [15]. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2015). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1), 107–124. Doi: 10.1111/exsy.12135
- [16]. Paterson, N. D. (2017). Predictors of first year retention rates at the university of the West Indies, Jamaica. *International Journal of Educational Development*, 55, 63–68. Doi: 10.1016/j.ijedudev.2017.06.001
- [17]. Porter, O. F. (2020). *Undergraduate Completion and Persistence at Four-Year Colleges and Universities: Completers, Persisters, Stopouts, and Dropouts*. National Institute of Independent Colleges and Universities
- [18]. Provencher, A., & Kassel, R. (2019). High-impact practices and sophomore retention: Examining the effects of selection bias. *Journal of College Student Retention: Research, Theory & Practice*, 21(2), 221–241. Doi: 10.1177/1521025117697728
- [19]. R Core Team. (2017). *R: A language and environment for statistical computing*. R- project. Retrieved from: <https://www.R-project.org/> [accessed: 01 September 2023].
- [20]. Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, 33, 1–39. Doi: 10.1007/s10462-009-9124-7
- [21]. Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68, 458–472. Doi: 10.1016/j.compedu.2013.06.009
- [22]. Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. *Medical Teacher*, 39(7), 757–767. Doi: 10.1080/0142159x.2017.1309376
- [23]. Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4(1). Doi: 10.1186/1755-8794-4-31
- [24]. Xgboost. (n.d.). *Introduction to Boosted Trees — xgboost 1.5.1 documentation*. Xgboost. Retrieved from: <https://xgboost.readthedocs.io/en/stable/tutorials/model.html> [accessed: 02 September 2023].