MERAL Publications and Solutions Pvt. Ltd.

Original Article

Prediction of university dropouts through random forest-based models

Fred Torres-Cruz[1*], Elqui Yeye Pari-Condori[2], Ernesto Nayer Tumi-Figueroa[2], Leonel Coyla-Idme[2], Jose Tito-Lipa[2], Leonid Aleman Gonzalez[2], Alfredo Tumi-Figueroa[2,3]

[1]Computer Science Research Institute, Universidad Nacional del Altiplano de Puno, Puno, Peru. [2]Faculty of Statistical Engineering and Computer Science, Computer Science Research Institute, National University of the Altiplano of Puno, Puno, Peru. [3]Faculty of Human Medicine, Universidad Nacional del Altiplano de Puno, Puno, Peru.

Correspondence: Fred Torres-Cruz, Computer Science Research Institute, Universidad Nacional del Altiplano de Puno, Puno, Peru. ftorres@unap.edu.pe

ABSTRACT

This study presents a solution for predicting university dropout rates, leveraging advanced digital technologies and the Random Forest algorithm. The model was developed using key academic variables, such as year of enrollment, program of study, semester attended, and academic performance, represented by the grade point average (GPA). A dropout threshold was established for students whose GPA fell below 11. The dataset was partitioned into 70% for training and 30% for testing, yielding an overall accuracy of 85.9%. Feature importance analysis identified semester and year of enrollment as the most influential factors in predicting dropout. While the model demonstrated a 91% accuracy in identifying students unlikely to drop out, its predictive capacity for students at risk of dropping out was limited to 52%. This approach constitutes a significant advancement in the implementation of digital technologies in education, enabling proactive strategies to improve student retention through data-driven predictive interventions.

Keywords: University dropout, Prediction, Random forest, Academic performance, Retention

Introduction

This work is proposed as a response to the growing concern about university dropout, a phenomenon with significant implications for educational institutions and social and economic development [1]. In this context, the fundamental purpose of this study was to develop a predictive model capable of accurately identifying students at risk of dropping out, using an approach based on the analysis of academic data through the Random Forest algorithm [2]. This method was selected for its ability to handle large volumes of heterogeneous data and its efficiency in interpreting complex variables. Thus, the research aims to provide a robust predictive tool that facilitates the implementation of proactive and personalized interventions to improve student retention and reduce university dropout rates [3, 4].

Previous studies on university dropout have addressed the issue from diverse perspectives, encompassing both socioeconomic and academic factors, revealing a multiplicity of variables influencing the phenomenon. For instance, the research of Acero (2019) and Ahmed and Khan (2019) laid the groundwork by highlighting the importance of institutional commitment and social integration factors in student retention—elements that have been complemented in recent years with the development of quantitative predictive models [5-7]. More recent studies have employed machine learning algorithms such as logistic regression and neural networks, demonstrating promising results in predictive accuracy. However, the use of Random Forest for dropout prediction is still emerging, offering advantages in terms of variable interpretability and robustness against noisy data, aspects this study explores and optimizes [8, 9].

In terms of objectives, this study seeks to achieve three specific and measurable goals: firstly, to construct a Random Forest-based prediction model capable of achieving high accuracy in

Access this article online	
Website: www.japer.in	**E-ISSN:** 2249-3379

How to cite this article: Torres-Cruz F, Pari-Condori EY, Tumi-Figueroa EN, Coyla-Idme L, Tito-Lipa J, Gonzalez LA, et al. Prediction of university dropouts through random forest-based models. J Adv Pharm Educ Res. 2025;15(1):78-83. https://doi.org/10.51847/PFb18QB60j

classifying students; secondly, to evaluate the relative importance of specific academic variables, such as the engineering school to which a student belongs, year of enrollment, semester attended, and GPA, in determining dropout risk; and finally, to provide an interpretative tool enabling institutions to identify and understand dropout patterns across different student cohorts, facilitating the design of tailored interventions. These objectives ensure a pragmatic and measurable approach that can be integrated into educational policy decision-making processes.

The justification for this research lies in its innovative approach, applying a machine learning algorithm—Random Forest—in the context of university dropout prediction, a methodology that has demonstrated consistent results in other domains but remains underexplored in this specific field [10-12]. Furthermore, the selection of key academic variables allows the analysis to focus on aspects directly linked to student performance, avoiding external factors that are harder to modify, such as socioeconomic conditions [5]. This makes the research an original and practical tool for educational institutions, aimed at improving the efficiency of retention strategies through a purely academic lens, with the potential to be implemented in digital student monitoring platforms [13, 14].

This study not only seeks to advance theoretical knowledge on factors associated with university dropout but also to provide an applicable and scalable tool that enables educational institutions to anticipate dropout and, ultimately, improve retention rates. The innovation of using Random Forest for the selection and analysis of academic variables represents a significant contribution to the field of education, with a technological approach that facilitates the translation of these findings into decision-support platforms, thereby enabling personalized and proactive interventions in academic settings [15].

Materials and Methods

The methodological design of this research focuses on analyzing the phenomenon of university dropout from a quantitative perspective, aimed at identifying predictive patterns in students' academic data. This phenomenon is approached as a critical indicator of academic efficiency in higher education institutions, as student retention is directly associated with educational quality and the social impact of universities. The adopted approach seeks not merely to view dropout as a simple statistic but to understand it through the lens of its academic predictors, enabling the design of data-driven, informed strategies to intervene at the early stages of the educational process.

The study is classified as quantitative, explanatory, and deductive, as its objective is to establish causal and correlative relationships between specific academic variables and dropout risk. The explanatory nature of the research is based on its intention to clarify the internal factors contributing to university dropout, which are also manipulable by institutions. Additionally, this is a cross-sectional study, given that the data spans multiple cohorts over a single academic period. This approach allows the evaluation of the phenomenon within a specific temporal context without experimental intervention.

Given the predictive focus, the central hypothesis of the research posits that "academic variables such as program of study, semester attended, and year of enrollment have a significant influence on the probability of university dropout, and these variables can be used to develop a robust predictive model based on the Random Forest algorithm." This hypothesis facilitates the exploration of academic data as early indicators of dropout, providing a solid foundation for proactive interventions in educational administration.

The study's scope is centered on the National University of the Altiplano of Puno, Peru, a large institution with an educational population of approximately 18,000 enrolled students. This diversity in academic programs and demographics makes the findings broadly applicable while maintaining controlled generalization. The study population includes all undergraduate engineering students at this institution over three years (2020, 2021, 2022), encompassing various engineering disciplines and levels of academic progress. This temporal span allows the analysis of dropout behavior patterns potentially influenced by changes in internal policies or external contexts, such as shifts in the educational or economic system.

For sample selection, 249,043 records of enrollments and grades from 2020, 2021, and 2022 were included. Although no specific sampling method was employed, this sample size ensures an adequate representation of the university's diverse student population. Furthermore, the large dataset provides sufficient statistical power for regression analyses and feature importance evaluations applicable to the Random Forest algorithm.

The variables analyzed were selected based on their theoretical and practical relevance to dropout prediction, aligning with the study's objectives. The independent variables include year of enrollment, program of study, semester attended, and academic performance (measured via GPA). The dependent variable is the probability of dropout, determined using a performance threshold, with a GPA below 11 indicating a risk of dropout. These variables not only allow robust quantitative measurement but also have a strong theoretical foundation in previous dropout studies, which link academic performance and curriculum progression to student attrition. Additionally, course grouping by student was applied to ensure accurate dropout identification, which enabled the analysis of the following datasets effectively.

Table 1. Análisis luego de la Segmentación

Variable	X	SD
Grades	12.79	3.34
	n	**%**
Dropout (No)	18,979	83%
Dropout (Yes)	3,833	17%

To implement the predictive model, the Random Forest algorithm was used, chosen for its effectiveness in data classification and its ability to handle variables with non-linear relationships, which is particularly relevant given the

heterogeneous nature of academic data. This algorithm also provides advantages in interpreting feature importance, which is crucial for identifying the factors that carry the most weight in predicting dropout. To ensure the model's validity and reliability, the data was partitioned into a training set (70%) and a testing set (30%), adhering to best practices in predictive modeling.

The model's accuracy was assessed using standard metrics, such as overall accuracy, sensitivity (the ability to correctly identify students at risk of dropping out), and specificity (the ability to identify students who do not drop out). These metrics allowed for evaluating the model's robustness and applicability in real-world scenarios. Variable importance was quantified based on each feature's contribution to the prediction, with semester attended and year of enrollment emerging as the most influential predictors. The selection of techniques, metrics, and procedures in this methodological design aligns with the need to develop a highly reliable predictive tool that can be effectively implemented in educational contexts.

The fieldwork for this research was conducted through a systematic strategy for collecting academic and demographic data from the academic management systems of the National University of the Altiplano. Data collection was carried out over a specific period, from January 2020 to July 2023, enabling the creation of a comprehensive and representative dataset of students enrolled across various semesters and study programs. To ensure the completeness and accuracy of the data, multi-stage verification and cleaning processes were performed, including techniques such as deduplication, imputation, and grouping, as well as the removal of incomplete or inconsistent entries. This process ensured an optimal level of data quality and mitigated biases stemming from recording errors, thus guaranteeing the validity of the dataset for subsequent analysis.

The data analysis employed data mining and machine learning techniques, focusing on the Random Forest algorithm due to its high classification capability and effectiveness in handling complex data with non-linear relationships. This technique was specifically chosen for its suitability in identifying patterns within datasets containing multiple predictive variables, as demonstrated in other studies, which is crucial for modeling dropout probability. The analysis was conducted in multiple phases: first, a training phase was performed using 70% of the dataset; then, the model was applied to the remaining 30% testing set to evaluate its predictive performance. Cross-validation and accuracy metrics were employed to optimize hyperparameters, ensuring the model's robustness and generalizability to new data, as outlined in **Figure 1**.

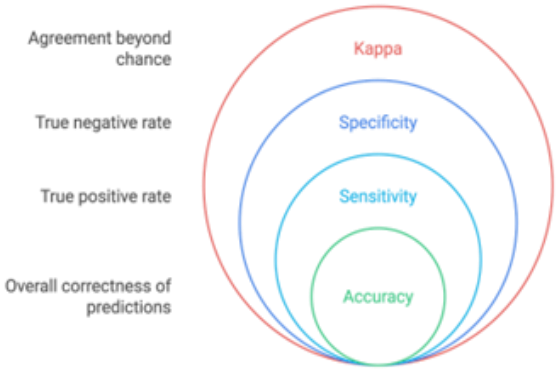

Figure 1. Metrics of Evaluation

The type of data analysis employed included predictive performance metrics such as accuracy, sensitivity, and specificity to evaluate the model's effectiveness in identifying students at risk of dropout. The model achieved an overall accuracy of 85.9%, with a sensitivity of 52% for detecting at-risk cases and a specificity of 91% for identifying students who continued their studies. These metrics were critical for assessing the model's capability in a real-world context, as accuracy alone is insufficient to interpret predictive validity in educational intervention applications. To determine the importance of each variable, the "Mean Decrease in Impurity" (MDI) metric was used, providing a quantitative assessment of each predictor's contribution to reducing uncertainty in classification.

Results and Discussion

The findings of this study reveal significant patterns in the prediction of university dropout through an exhaustive analysis using the Random Forest model. The model achieved an overall accuracy of 85.9%, demonstrating robust predictive capacity in distinguishing students likely to persist in their studies from those at risk of dropping out. Performance was evaluated using metrics of accuracy, sensitivity (recall), and specificity. While the model exhibited a specificity of 91%, indicating high efficiency in classifying persistent students, its sensitivity was limited to 52% for identifying at-risk students. This discrepancy highlights the model's strength in recognizing non-dropouts but also its relatively lower performance in detecting students at risk.

The analysis of feature importance for the predictive variables underscored that semester attended and year of enrollment are significant factors influencing dropout prediction. The contributions of these variables were quantified using the Mean Decrease in Impurity (MDI) metric, which measures each predictor's role in reducing node impurity within the decision trees of the Random Forest. The semester attended exhibited the highest MDI value (MDI_semester = 0.328), followed by a year of enrollment (MDI_ingreso = 0.296), program of study (MDI_carrera = 0.214), and GPA (MDI_promedio = 0.162). These results confirm that academic progress, as reflected in the semester and year of enrollment, plays a pivotal role in determining dropout probability.

To further evaluate model performance, a Receiver Operating Characteristic (ROC) curve was constructed, and the Area Under the Curve (AUC) was calculated. The AUC value of 0.77 indicates reasonable discrimination between students who drop out and those who persist. The ROC curve, derived from sensitivity and specificity values at various classification thresholds, confirmed the model's adequate yet improvable performance, particularly in terms of sensitivity. The lower sensitivity in identifying at-risk students suggests opportunities for model optimization through techniques such as class imbalance adjustment or ensemble algorithms designed to enhance sensitivity without compromising specificity.

The variable "GPA" also exhibited a strong correlation with dropout probability. Students with GPAs below 11 showed a marked tendency to drop out, validating the threshold established for classifying at-risk students. Notably, prediction accuracy increased for students with low GPAs, suggesting that this variable is an effective predictor, particularly when considered alongside other academic factors such as semester attended. This finding underscores the importance of targeted intervention strategies for students with poor academic performance. Prioritizing such students through proactive measures could significantly mitigate dropout rates, reinforcing the utility of GPA as a key indicator for early intervention.

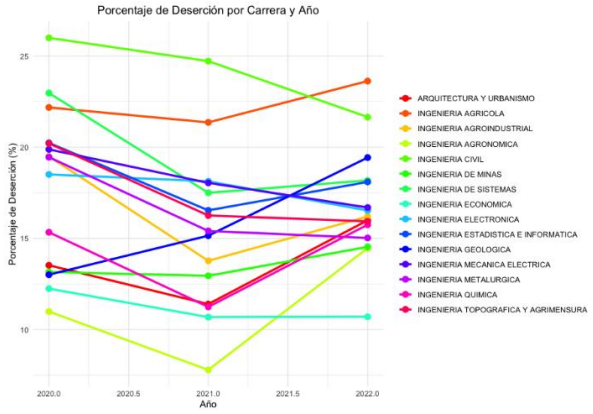

Figure 2. Dropout by Career

Additionally, the results of the complementary logistic regression applied to key variables confirmed the influence of semester attended (B_semester = 0.452, $p < 0.001$) and year of enrollment (B_ingreso = 0.376, $p < 0.001$) as significant predictors. This additional analysis was conducted to assess the stability of these variables within a statistical framework, and its findings reinforce the robustness of the Random Forest model by identifying the same influential predictors. The alignment between both models provides theoretical and practical support for the importance of these variables in the dropout phenomenon, underscoring the need to focus interventions during these critical periods.

In terms of classification errors, the analysis revealed that the model has a relatively high false-negative rate compared to false positives. This means there is a greater likelihood of failing to identify at-risk students than erroneously classifying students as at-risk. The false-negative rate was 48%, while the false-positive rate was 9%. This error pattern suggests that the model prioritizes ensuring that students flagged as at risk are indeed critical cases, at the expense of potentially overlooking some dropout cases. This trade-off should be carefully considered in the design of preventive policies.

Category-specific accuracy demonstrated that certain programs exhibit higher dropout rates than others, influenced by factors such as semester attendance and academic performance. This finding supports the implementation of tailored academic support strategies that address the unique characteristics of each program, thereby improving the model's accuracy for future applications across diverse academic contexts.

The results also revealed an accumulated risk trend by semester, showing that students in the early semesters (1st to 3rd) face a higher dropout rate than those in advanced semesters. This phenomenon was quantified using a Kaplan–Meier curve applied to semester progression, which displayed a significant decrease in the probability of dropout as students advanced in their academic careers. This trend underscores the critical importance of early and targeted interventions designed to mitigate the risk of attrition during the initial academic cycles when students are most vulnerable.

The findings provide a solid empirical foundation for implementing a predictive monitoring system within the university's digital platforms, enabling the development of personalized interventions. Although the model's sensitivity is limited, its ability to identify at-risk students offers significant value as a strategic tool for academic retention efforts. This study suggests that by optimizing the model's sensitivity and fine-tuning classification parameters, it is possible to enhance dropout prediction accuracy and design more effective preventive policies.

Conclusion

The findings of this research demonstrate that university dropout can be predicted with considerable accuracy using advanced machine learning algorithms, specifically the Random Forest model [16]. The constructed model achieved an overall accuracy of 85.9% and a specificity of 91%, confirming its ability to identify students with a low probability of dropping out. However, its sensitivity of 52% indicates moderate effectiveness in identifying students at risk. Semester attended and year of enrollment emerged as the most influential variables, highlighting academic progress as a critical predictive factor in higher education attrition. This finding reinforces the validity of using academic indicators as central features in predictive dropout models, as opposed to external contextual variables [17].

The initial hypothesis of this study posited that specific academic variables, particularly semester attended and year of enrollment, would have a significant influence on the likelihood of dropout and that a Random Forest-based model could effectively capture these relationships. The results strongly support this hypothesis, demonstrating that these variables are not only robust predictors

but also carry considerable weight in classifying at-risk students. The implementation of the model validated the notion that exclusively academic data can be effective for generating actionable predictions without relying on variables that are difficult to intervene upon, such as students' socioeconomic characteristics.

A comparative analysis of the results with prior studies reveals that while earlier research has employed demographic and socioeconomic variables to complement academic performance, the findings of this study suggest that a purely academic focus can be highly effective for predicting dropout [18]. This represents a significant methodological advancement, as it allows intervention efforts to be concentrated on factors directly controllable by institutions [19]. Compared to models based on logistic regression or neural networks, the Random Forest algorithm not only offers competitive accuracy but also provides an advantage in interpretability by identifying the most influential variables through techniques such as feature importance analysis [20].

Despite these promising results, the relatively low sensitivity in identifying at-risk students highlights a limitation in the model's ability to detect certain dropout cases, potentially underestimating the true scope of the problem. This suggests the need to explore additional machine learning approaches or class imbalance adjustment methods that could enhance detection capability without compromising the accuracy of persistent student classification [5]. Furthermore, these findings open the door to combining Random Forest with ensemble techniques, such as boosting or incorporating longitudinal tracking data to provide a more comprehensive view of factors evolving over the academic cycle.

This study provides a solid foundation for implementing predictive models of university dropout focused on academic performance, demonstrating that internal variables within the educational process are sufficient to generate early warnings for at-risk students [14, 21, 22]. The novelty of this approach lies in its ability to simplify interventions by relying exclusively on academic data, making it particularly suitable for integration into student management systems [20, 23]. These findings, along with the flexibility and scalability of the Random Forest model, make a significant contribution to the field of academic retention, offering an analytical and actionable framework that can be adapted and optimized across various educational institutions to effectively and proactively address the issue of dropout.

Acknowledgments: We gratefully acknowledge the university for its unwavering support and the students for their valued participation in this endeavor.

Conflict of interest: None

Financial support: Financial support for this study was provided by the Universidad Nacional del Altiplano de Puno. The project was funded through resources from FEDU.

Ethics statement: Informed consent was obtained from all participants prior to their involvement. Each participant received comprehensive information about the study's aims, procedures, potential risks, and benefits, and was assured of confidentiality and the right to withdraw at any time.

References

1. Vargas LP, Nuñez EJ, Ruge IA, López FR. Influential factors in the desertion of electronic engineering students from UPTC admitted in 2015. In2021 International Symposium on Accreditation of Engineering and Computing Education (ICACIT) 2021 (pp. 1-5). IEEE. doi:10.1109/ICACIT53544.2021.9612503
2. Sharma N, Sharma M, Garg U. Predicting academic performance of students using machine learning models. In2023 International Conference on Artificial Intelligence and Smart Communication (AISC) 2023 (pp. 1058-63). IEEE. doi:10.1109/AISC56616.2023.10085214
3. Saavedra-Acuna C, Quezada-Espinoza M, Correa DA. Analyzing attrition: predictive model of dropout causes among engineering students. In2024 ASEE Annual Conference & Exposition 2024.
4. Al-Johani K, Jamal BT, Hassan M, Al-Sebaei MO. Knowledge and attitude of Dental students towards Medical emergencies at King Abdulaziz University, Jeddah, Saudi Arabia. Ann Dent Spec. 2022;10(1):137-40. doi:10.51847/yRxgh18NN1
5. López AE, Achury JC, Morales JC. University dropout: a prediction model for an engineering program in bogotá, Colombia. InProceedings of the 8th Research in Engineering Education Symposium, REES 2019-Making Connections 2019 (pp. 483-90).
6. Ahmed SA, Khan SI. A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective. In2019 10th international conference on computing, communication and networking technologies (ICCCNT) 2019 (pp. 1-6). IEEE. doi:10.1109/ICCCNT45670.2019.8944511
7. Shchuchka RV, Kravchenko VA, Zakharov VL. Biological efficiency of the application of herbicides on spring barley in the chernozem region. Entomol Appl Sci Lett. 2021;8(3):8-14. doi:10.51847/h7kM3hIIuB
8. Amaya AB, Barrera HC, Manrique R. Analysis of machine learning models for academic performance prediction. In International Conference on Intelligent Tutoring Systems 2024 (pp. 150-161). Cham: Springer Nature Switzerland. doi:10.1007/978-3-031-63031-6_13
9. Almulhim FA, Alshahrani MM, Hakami AM, Shammaa AM, Aljehaiman TA, Alsaihati AM, et al. Review on pneumothorax diagnostic and management approach in emergency department. Int J Pharm Res Allied Scie. 2022;11(1):35-9. doi:10.51847/597Cjlr708
10. Alcauter I, Martinez-Villasenor L, Ponce H. Explaining factors of student attrition at higher education.

Computación y Sistemas. 2023;27(4):929-40. doi:10.13053/CyS-27-4-4776

11. Albalwei HS, Ahmed NF, Albalawi NM, Albalawi SS, Al-Enazi NH. Violence against health care workers of pediatric departments in Saudi Arabia: systematic review. Arch Pharm Pract. 2021;12(1):79-84. doi:10.51847/RJ683KlDZS
12. Enwa S, Ogisi OD, Ewuzie PO. Gender role and effects on climate change adaptation practices among vegetable farmers in delta central zone. World J Environ Biosci. 2024;13(1):22-9. doi:10.51847/hJorfK74GJ
13. Albalawi Y, Buckley J, Nikolov NS. Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media. J Big Data. 2021;8(1):95. doi:10.1186/s40537-021-00488-w
14. Vega KA, Tejeda S. Exploring the transformative potential of machine learning in engineering and STEM education: a comprehensive study on enhancing cognitive levels and learning insights. In2023 World Engineering Education Forum-Global Engineering Deans Council (WEEF-GEDC) 2023 Oct 23 (pp. 1-9). IEEE. doi:10.1109/WEEF-GEDC59520.2023.10343631
15. Torres-Cruz F, Yucra-Mamani YJ. Artificial intelligence techniques in assessment of virtual education by university students. Int Humanit Rev. 2022;11(4):16. doi:10.37467/revhuman.v11.3853
16. Canto NG, de Oliveira MA, Veroneze GD. Supervised learning applied to graduation forecast of industrial engineering students. Eur J Educ Res. 2022;11(1):325-37. doi:10.12973/eu-jer.11.1.325
17. Canqui-Flores B, Mendoza-Mollocondo CI. Torres-Cruz, F, Fuentes-López J, Gómez-Campos R, Viveros-Flores A, et al. Validity, reliability and scale to measure the self-perception of academic stress of university students. Gac Med Bilbao. 2019;116(4):158-65.
18. Dávila G, Haro J, González-Eras A, Vivanco OR, Coronel DG. Student dropout prediction in high education, using machine learning and deep learning models: case of ecuadorian university. In2023 International Conference on Computational Science and Computational Intelligence (CSCI) 2023 (pp. 1677-1684). IEEE. doi:10.1109/CSCI62032.2023.00277
19. Lu X, Tong J, Xia S. Entity relationship extraction from Chinese electronic medical records based on feature augmentation and cascade binary tagging framework. Math Biosci Eng. 2024;21(1):1342-55. doi:10.3934/mbe.2024058
20. Bussaman S, Nasa-Ngium P, Sararat T, Nuankaew WS, Nuankaew P. Influence analytics model of the general education courses toward the academic achievement of rajabhat university students using data mining techniques. InWorld Conference on Information Systems for Business Management 2023 (pp. 117-29). Singapore: Springer Nature Singapore. doi:10.1007/978-981-99-8612-5_10
21. Tocto Inga PM, Huamaní Huamaní GT, Zuloaga Rotta LA. Machine learning application in university management: classification model Dropping out of engineering students in Peru. LACCEI. 2023;1(8).
22. Arakelyan LA, Kamentseva PG, Mashakova AD, Kolesnichenko VV, Karoli II, Voropaev VV. Preparation of a new enterosorbent bentorb and determination of its toxicological properties. Pharmacophore. 2024;15(2):105-12. doi:10.51847/d4HrIDWggY
23. Booth GJ, Ross B, Cronin WA, McElrath A, Cyr KL, Hodgson JA, et al. Competency-based assessments: leveraging artificial intelligence to predict subcompetency content. Acad Med. 2023;98(4):497-504. doi:10.1097/ACM.0000000000005115