Advancing Education: Hybrid Recommendation Systems for Best-Fit Student Domain Matching

Sarra Aouadi[1,*], Toufik Marir[1], and Mohammed Lamine Kherfi[2,3]

[1] ReLa(CS)[2] Laboratory,University of Oum El Bouaghi,Oum El Bouaghi, Algeria
{aouadi.sara,marir.toufik}@univ-oeb.dz
[2] LAMIA Laboratory, University of Quebec in Trois-Rivières, Trois-Rivières, Canada
Kherfi@uqtr.ca
[3] Department of Computer Science, University of Kasdi Merbah, Ouargla 30000, Algeria
[*] Corresponding author

Abstract. Universities around the world are concerned with the student dropout phenomenon, which is particularly prevalent in the early years. Research indicates that the main reason for early dropout is the wrong choice of academic study domain. In this work, we have tried to provide decision-making support to the new students to help them choose the path that best suits their abilities and skills.From a conceptual perspective, we propose a hybrid recommendation system that integrates machine learning algorithms and collaborative filtering techniques to address real-world educational big data.From a practical standpoint, this system utilizes the machine learning model to identify the academic domain in which a student is most likely to succeed. Subsequently, collaborative filtering is applied to utilize the top 20% of similar students to estimate potential success rates within the predicted domain.Our approach introduces several significant innovations compared to existing methods, demonstrating improved prediction accuracy and offering the potential to positively impact academic success rates

Keywords: Dropout, Academic domain, Machine learning model, Recommendation system

1 Introduction

Graduating from high school allows students to automatically join the higher education system and start a new journey. Even though universities around the world aim to improve the student's future life by developing his personality and ameliorating his technical skills, Harvard University and the Asian Development Bank recently reported that university degree holders represent only 6.7 % of the world's population [1]. This low indicator reflects that not all enrolled students have been able to pass the entire academic program. United States College graduation statistics indicate that of the total university graduates, almost half hold a bachelor's degree, with 23.5% of bachelor's degree holders taking longer

C. A. Kerrache et al. (eds.), *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*, Advances in Intelligent Systems Research 184,
https://doi.org/10.2991/978-94-6463-496-9_6

than the stipulated period to complete. And 32.9% of dropped-out students had no academic degree [2]. South African universities have also been hit by this phenomenon, with a dropout ratio exceeding 50%, specifically in first-year students [3]. Also, in France, the Ministry of Higher Education and Research reports that there are significant increases in the failure rate in the first cycle, especially in the first year of a bachelor's degree [4], and this is most likely due to university orientation policies that sometimes lead students to make "wrong" choices. Undeniably, Dropout has become a concern of many universities all over the map because dropping out of the higher education system without obtaining at least a first degree has a negative impact not only on the individual and university institutions but also on society.

Considering Algeria's case, the higher education sector receives annually thousands of baccalaureate degree holders or any equivalent diploma for the pursuit of studies in the BMD (Bachelor, Master, and Doctorate) system. Despite the efforts to improve the academic sector's outcomes, the situation is similar to the previously mentioned countries. As the Algerian Ministry of Higher Education and Scientific Research reported, there is a growing concern about student success in the first year. The gap between the number of those enrolled in the first year and those in the second year cannot be ignored, which is somehow related to the large number of repeaters, re-orientations (change the educational field), and those who leave higher education entirely without any academic achievement. These difficulties are mainly related to a poor orientation where students are sent to fields that neither interest nor suit them. That is why the student must make his decision by considering many factors to ensure his success, not definitely, but at least sure that he has chosen what suits him best. The main challenges for higher education institutions are the selection and assignment of a suitable student for enrolment in a particular field, and all enrolled students must have achieved their academic degree within the given delay.

In this work, we consider the orientation problem of students as a recommendation problem. Consequently, we propose a hybrid recommendation system based on machine learning algorithms and collaborative filtering. The remainder of this paper is organized as follows: related works are presented in section 2 then, in section 3 we explain our approach. Finally, the conclusion and future works are presented in section 4.

2 Related Work

In this educational context, Educational Data Mining (EDM) has emerged as a notable approach, facilitating the examination of large educational datasets and the extraction of hidden information where the intersection of artificial intelligence (AI) models with the prediction of student behavior has garnered significant attention [5].

Studies have effectively harnessed machine learning algorithms and data mining techniques to analyze diverse educational data, offering valuable insights into learning patterns and engagement levels. The prior research by Sekeroglu et

al. [6] has significantly contributed to this field by advocating for applying five machine-learning algorithms in predicting and classifying student performance. Their practical applications on datasets from school reports and questionnaires demonstrated the effectiveness of these algorithms. Experimenting with various data selection methods and algorithms can enhance the performance.

Furthermore, a comparative study affirmed the efficacy of regression by random forest in predicting performance [7].The effectiveness of Random Forest stands out more in predicting learning outcomes using a machine learning algorithm based on online learning behavior data in blended courses [8]. The researchers investigate how online learning behaviors, including engagement, participation, and interaction patterns, can predict students' academic success. Their research aims to offer insights into effective online learning strategies and support systems for students in blended courses. However, the system's reliance solely on online learning behavior data may not capture all factors influencing learning outcomes. Additionally, the predictive model's generalizability might be limited to specific course contexts or student populations. Further validation and refinement of the model in real-world educational settings are needed.

Previous studies have explored the use of decision trees in various contexts, such as detecting students with high academic potential and predicting academic performance. Decision trees are favored for their ability to simplify models, making them more interpretable by experts. However, limitations arise due to the absence of critical variables like gender, scholarships, and financial aid, impacting the consistency of classification algorithms. Evaluating equity and discrimination in predictive models remains a challenge across these studies [9]. For instance, a prior study achieved high accuracy with a decision tree classifier to predict students' academic performanc but overlooked other important performance metrics [10]. Similarly, Hussain et al. (2023) employed a genetic algorithm-based decision tree regression to predict exam marks and forecast grades for students in Pakistan [11]. Their model demonstrated the effectiveness of educational data mining in accurately anticipating student achievement levels, enabling proactive identification of those in need of intervention.

Moreover, Recent analyses have explored the application of classification algorithms in the context of academic track selection in Saudi universities. One study investigated the efficacy of the Logistic Regression algorithm in developing predictive models based on Grade Point Average (GPA), guiding students in selecting suitable academic tracks based on preparatory year performance [12]. Notably, the current focus in academic research is on comparing machine learning models for predicting student success, but these studies often need to pay more attention to critical demographic factors. This omission limits the accuracy and completeness of predictive models.

Artificial neural networks have also played a pivotal role in predicting student behavior. A model based on convolutional neural networks was proposed for student performance prediction, highlighting the close relationship between the learning process and student behavior [13]. Similarly, a deep neural network model was utilized to predict student's ability to pass future courses based

on previous performance, providing valuable insights for educational decision-making [14]. Regardless, their proposed deep learning algorithm can be complex for computers to understand, especially when dealing with large and complex datasets. Researchers showcased the effectiveness of predictive models based on deep learning techniques in forecasting academic performance [15]. At the same time, natural language processing was employed to predict student engagement by analyzing online forum discussions [16].
In recommendation systems, notably, their application has expanded beyond e-commerce to include education. A web-based decision support system named orieB was developed, combining content-based and collaborative filtering recommender system techniques to assist secondary students in optimal baccalaureate affiliation [17]. Also, a hybrid recommendation system was employed for an e-orientation platform as an assistant for advisors in recommending the study path appropriate to a student's profile. Student's data were collected from questionnaires and tests [18]. One limitation of this work is its inability to ascertain its suitability for handling a large number of users. TIMONEL, an advisory tool, guides academic higher education students based on a multi-agent intelligent system, emphasizing careful consideration of recommendations by advisors before presentation to students [19]. Yet, a significant constraint limitation in the current state of the art is the need to integrate demographic features, which is essential for achieving a more nuanced understanding of student behavior and enhancing predictive accuracy.

All the aforementioned works have achieved improvements in specific areas such as the volume of collected data, data processing methods, accuracy, model interpretability..etc. Given the sensitivity and growing complexity of the student dropout issue in the higher education sector, it is imperative to develop a new system that integrates the highest number of quality features and characteristics. This motivates our research, detailed in the following sections, to create a new hybride recomendation system that effectively addresses student dropout.

3 The Proposed Approach

This research introduces a novel hybrid recommendation system designed to assist students in selecting optimal academic choices within Algerian universities. The system integrates machine learning algorithms and collaborative filtering techniques to assess the suitability of a student's initial academic decision and provide recommendations for more appropriate academic domains. By considering the predictive domain assignment derived from machine learning models and the collective experiences of similar students, this hybrid recommendation system aims to enhance academic advising and promote student success. Figure 1 shows the overall process of the proposed approach. Developing our approach involves several steps, such as collecting data, data preprocessing, selecting the machine learning model, applying the recommendation technique, evaluating the results, and presenting the final recommendations. The remainder of this section is devoted to detailing these steps.

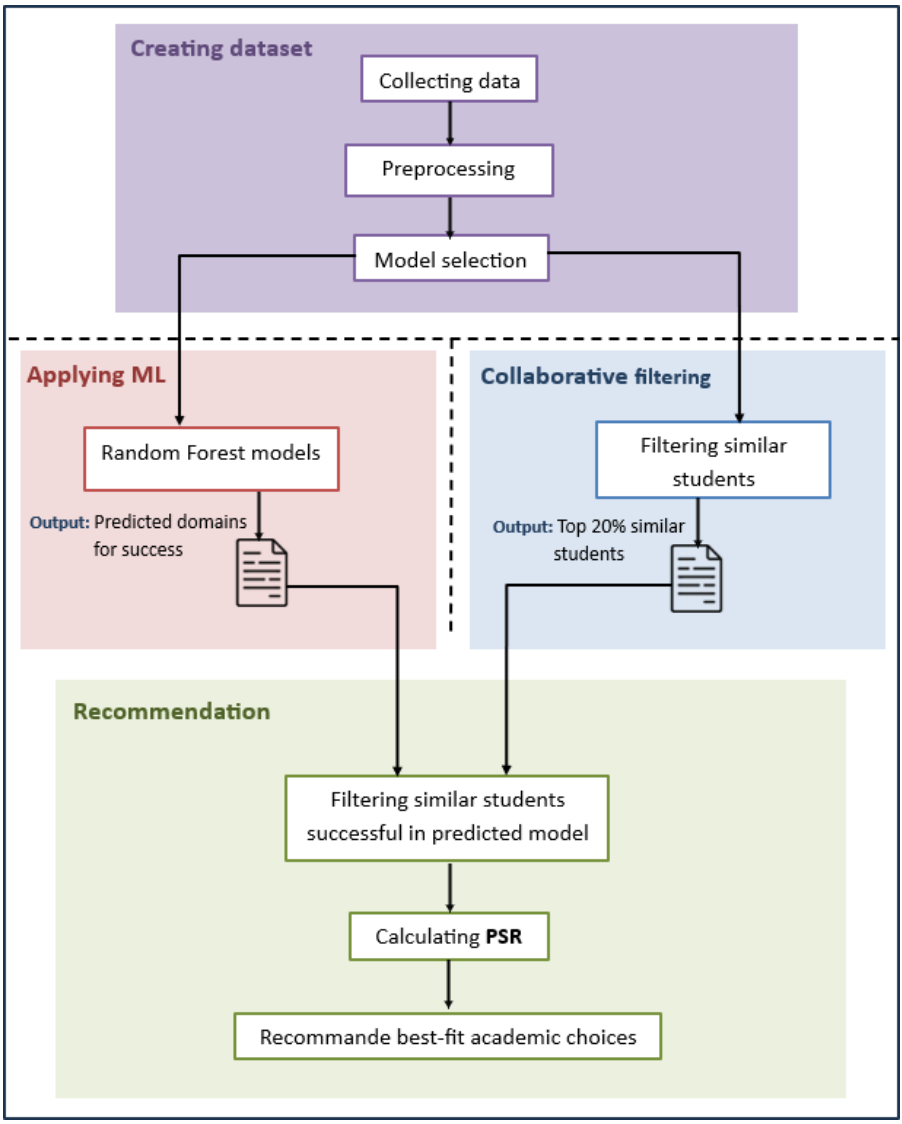

Fig. 1. The overall process of the proposed approach.

3.1 Data collection

Thanks to the digitization policy adopted by the Algerian Ministry of Higher Education and Scientific Research, a large amount of digital data become available. Almost all the pedagogical and administrative processes, like admission and orientation processes, are facilitated through a dedicated platform known as "PROGRES" [20].
In our study, we focused on the academic results of the first year because it represents the critical academic step for most students. Hence, this research takes advantage of comprehensive data that includes information on more than 520,000 first-year students enrolled between 2021 and 2023. It is important to note that we have used, in addition, the academic data (high school graduation year, specialization, grades, university grades... etc.), demographic data (like gender, region family background, parents' income, parents' civil status) and administrative one (like registration date, student ID, level, institution... etc.). As we mentioned previously, the problem of university dropout is closely related not only to academic reasons but also to the demographic background, which has a significant effect.

3.2 Data preprocessing

Effective data preprocessing is crucial for extracting reliable information from the subsequent application of machine learning algorithms. The raw dataset is carefully processed to prepare it for modeling. Initial data cleaning involves handling missing values, smoothing outliers, correcting inconsistencies, and removing duplicates. Relevant transformations, including normalization, encoding categorical variables, and dimensionality reduction, are applied through feature selection. Additionally, the dataset is partitioned into suitable training and test sets for further analysis.

3.3 Selection of Machine Learning Algorithms

In this step, we aim to select the appropriate machine learning algorithms. The most effective way to do such selection is by comparing different techniques. Hence, in this work, we have compared three machine learning algorithms: Decision Trees, Random Forests, and Support Vector Machines.
We have used three evaluation metrics to compare the above algorithms: Recall, Precision, and Accuracy. Table 1 and Figure 2 give the results of this comparison. The Random Forests algorithm provides the best result, making it the best algorithm to apply in our domain.

The Random Forest model is trained on 80% of the dataset to predict the academic domains where a student will likely succeed. Features such as academic history, demographic data, and other relevant information are utilized to make this prediction.

Table 1. The result of comparing ML algorithms

	Decision Tree	Support Vector Machine	Random Forest
Recall	0.963	0.961	0.974
Percision	0.963	0.961	0.974
Accuracy	96.292	96.095	96.507

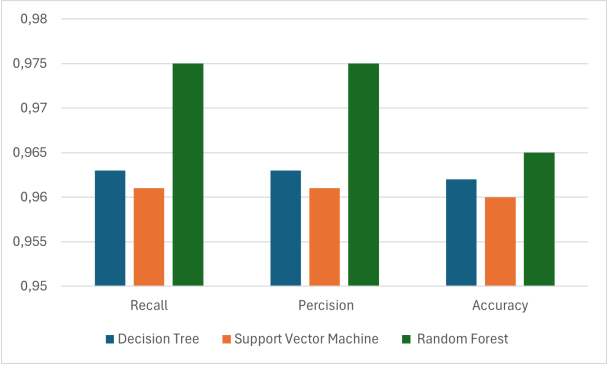

Fig. 2. Visualizing Performance: Comparing SVM, DT,and RF

3.4 Recommendation System Integration

Considering students' recommendations as a critical task, we think providing only the predicated domains for success is insufficient. As a novelty in our approach, we opted for a complementary step that compared the student with those most similar to him. Thanks to the collaborative recommendation filtering, the system selects the top 20% of students similar to the academic and demographic profile of the target student.

3.5 Calculating the Potential Success Rate (PSR)

Compared to the proposed approaches in this field, we introduced the similarity between students in our study. Consequently, we have combined the result of the Random Forest algorithm with the similarity of students to propose not only if the student will succeed or drop out in a choice, but also his Potential Success Rate (PSR) in each choice. This rate reflects the student's potential success in his predicted academic domain based on the collective experience of similar peers. Hence, this rate is calculated for each domain as the weighted average of the Grade Point Average (GPA) of similar students, where the weights are determined by the degree of similarity (S) and (n) represents the number of similar students in this domain (equation 1).

$$PSR = \frac{\sum_{i=1}^{n} GPAi * Si}{n} \tag{1}$$

A s a result, our approach gives several significant novelties compared to the proposed approaches in this domain. Firstly, our research stands out through its comprehensive inclusion of academic and demographic features, including geographical region and parental situation, which have been underexplored in prior works. This holistic consideration enhances the model's predictive capabilities, making it a valuable contribution to the field of student success prediction. Secondly, our approach's adopted machine learning techniques (Random Forest) give the best results compared to the other methods (Decision Tree and SVM), as shown. In addition, our approach provides the potential of success based on the collaborative recommendation technique. In fact, using only the Random Forest, we can conclude that a student will succeed, but upon analyzing the results of his similar peers, we find that their average grades are close to the failing threshold. In this case, this choice poses a risk for the student. On the other hand, the potential of success can be considered as a prediction of the student's average grade, allowing the best choice. Finally, our approach is based on real data from the Algerian context.

4 Conclusion

In this paper, we have highlighted the prevalence of the university dropout phenomenon around the world in general and in Algeria in particular, where the main reason for early dropout was the wrong choice of academic study domain. Then, we have discussed some of the relative research that worked on predicting at-risk students and suggesting a recommendation system that supports decision-making in the academic field. In our approach, we have proposed a hybrid recommendation system, combining machine learning and collaborative filtering, to address the critical issue of university dropout resulting from inappropriate academic domain choices. In the future, we intend to expand our research to cover not just the first year but the entire academic journey of students. Further analysis and consideration of other factors, such as computational complexity, interpretability, and scalability, may be necessary to decide on the most suitable algorithm for the given problem.

References

1. Diana Anghel. The 6.7% welcomes you. https://dailyillini.com/special-sections/graduation-guide/2021/04/25/the-6-7-welcomes-you/, 2021. Accessed: 2024-3-5.
2. Melanie M Hanson. College graduation statistics, June 12, 2022. https://educationdata.org/number-of-college-graduates.
3. K Moodley, M Van Wyk, D Tive, I Van Zyl, and J Cronje. Student retention in fully online programmes at two south african universities. *International Journal of Information and Education Technology*, (9), 2023.
4. Mathias Millet. L'«échec» des étudiants de premiers cycles dans l'enseignement supérieur en france. retours sur une notion ambiguë et descriptions empiriques. *Réussite, échec et abandon dans l'enseignement supérieur*, pages 69–88, 2012.

5. Cristobal Romero and Sebastian Ventura. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355, 2020.
6. Boran Sekeroglu, Kamil Dimililer, and Kubra Tuncal. Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th international conference on educational and information technology*, pages 7–11, 2019.
7. Ahajjam Tarik and Yousef Farhaoui. Recommender system for orientation student. In *Big Data and Networks Technologies 3*, pages 367–370. Springer, 2020.
8. Yangyang Luo, Xibin Han, and Chaoyang Zhang. Prediction of learning outcomes with a machine learning algorithm based on online learning behavior data in blended courses. *Asia Pacific Education Review*, 25(2):267–285, 2024.
9. JH Guanin-Fajardo, J Guaña-Moya, and J Casillas. Predicting academic success of college students using machine learning techniques. data 2024, 9, x, 2024.
10. Opeyemi Ojajuni, Foluso Ayeni, Olagunju Akodu, Femi Ekanoye, Samson Adewole, Timothy Ayo, Sanjay Misra, and Victor Mbarika. Predicting student academic performance using machine learning. In *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part IX 21*, pages 481–491. Springer, 2021.
11. Shah Hussain and Muhammad Qasim Khan. Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of data science*, 10(3):637–655, 2023.
12. Thamer Althubiti, Tarig M Ahmed, and Madini O Alassafi. Developing a predictive model for selecting academic track via gpa by using classification algorithms: Saudi universities as case study. *International Journal of Advanced Computer Science and Applications*, 14(5), 2023.
13. Jian Zong, Chaoran Cui, Yuling Ma, Li Yao, Meng Chen, and Yilong Yin. Behavior-driven student performance prediction with tri-branch convolutional neural network. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2353–2356, 2020.
14. Shuping Li and Taotang Liu. Performance prediction for higher education students using deep learning. *Complexity*, 2021:1–10, 2021.
15. A Smith, B Johnson, and C Brown. Deep learning for predicting student academic performance. *Journal of Educational Technology & Society*, 22(3):220–232, 2019.
16. M Jones and L Wang. Predicting student engagement in online discussions using natural language processing techniques. *Computers & Education*, 150, 2020.
17. Emilio J Castellano, Manuel J Barranco, and Luis Martínez. Academic orientation supported by hybrid intelligent decision support system. 2011.
18. Ghanjaoui Meryem, Khadija Douzi, and Salima Chantit. Toward an e-orientation platform: Using hybrid recommendation systems. In *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–6. IEEE, 2016.
19. Antonio Pantoja-Vallejo and Beatriz Berrios-Aguayo. Timonel: Recommendation system applied to the educational orientation of higher education students. In *International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning*, pages 14–26. Springer, 2019.
20. Ministère de l'Enseignement Supérieur et de la Recherche Scientifique. Progres, 2024.