Enhancing Educational Outcomes Through AI Powered Learning Strategy Recommendation System

Daminda Herath, Chanuka Dinuwan, Charith Ihalagedara, Thanuja Ambegoda
Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

***Abstract*—In order to develop intelligent learning recommendation systems, the work identifies the employment of artificial intelligence (AI) techniques, particularly in the educational data mining (EDM) field. The aggregation of such educational data into an efficient analytical system could also assist as an interesting means of education for the students. In fact, it could ultimately advance the direction of education. Sophisticated machine learning methods were employed to analyze various data types, including educational, socioeconomic, and demographic data, to predict student success. In this research, Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), CatBoost, and XGBoost algorithms were considered to build prediction models using a dataset encompassing a wide range of student traits. Robust evaluation metrics, including precision, recall, accuracy, and F1-score, were used to gauge model effectiveness. The results highlighted that RF was the best with accuracy, precision, and recall. Then, a rule engine was built to enhance the system by finding the most efficient learning tactics for students based on their expected future performance. The proposed AI-based personalized recommendation tool shows a substantial step towards enhancing educational decisions. This solution facilitates educators in creating student academic assistance interventions by offering individualized, data-driven learning strategies.**

Keywords*—Artificial intelligence; educational data mining; educational strategies; machine learning; personalized recommendation; student performance prediction

I. Introduction

The incorporation of online learning environments and traditional classroom settings within educational institutions is progressively supported by an increasing volume of educational data. These data-rich settings offer substantial opportunities but also pose considerable challenges as institutions struggle to effectively bind them to enhance educational outcomes [1]. A key challenge is the collection and analysis of massive datasets to produce meaningful insights that can significantly influence educational strategies [2].

The science of discovering novel patterns in vast volumes of data is known as data mining. Moreover, finding knowledge in educational data that could significantly influence research methods in the field of education is the goal of educational data mining, or EDM [3]. In the past few years, researchers in this field have been eager to get valuable and practical insights, especially about student performance [4]. The aim to assist educational establishments in making choices and improving their teaching strategies is the motivational drive. In addition to help students overcome obstacles to learn and enhance their academic performance [5].

Failure and dropout rates are still rising despite technical advancements in e-learning that have undoubtedly had an impact on student performance [6]. This situation demands for a complete examination of a variety of influencing factors, including demographics, personal experiences, academic data, institutional surroundings, socioeconomic standing, and individual self-efficacy [7]. In fact, understanding the relationship between these factors is critical for evolving effective learning strategies and supporting academic accomplishment [8].

Despite the availability of extensive educational data, institutions face challenges in effectively utilizing this information to predict student performance and develop personalized learning strategies. Current systems frequently do not account for the diverse socioeconomic, demographic, and academic factors that affect student success. Consequently, dropout rates remain high, and students do not receive the individualized support necessary to enhance their academic outcomes.

In this study, the research objectives are; (1) to identify the key demographic, socioeconomic, and academic factors that influence student performance in higher education; (2) To apply advanced data mining techniques and machine learning models to predict student outcomes, specifically focusing on graduation, dropout, and enrollment statuses; and (3) To design and implement an AI-powered learning strategy recommendation that delivers personalized educational interventions based on predicted student performance.

The main contributions and implications of this study include the identification of key factors influencing student performance, the development of a predictive model utilizing machine learning algorithms to forecast student outcomes, and the introduction of an AI-powered personalized recommendation system that suggests individualized learning strategies based on performance predictions. This system enables educators to provide tailored interventions, which may help reduce dropout rates and enhance academic outcomes, while also delivering actionable insights for institutional policies and resource allocation. Furthermore, the system serves as a practical tool for educators to intervene early in students' academic journeys, allowing students to benefit from personalized learning strategies.

All in all, the structure of this paper is as follows: literature review in Section II, research methodology in Section III, results and evaluation in Section IV, discussion in Section V and finally conclusion in Section VI.

II. Related Works

Numerous research studies are included with data mining techniques with classification and clustering [9], prediction modeling [10], and AI-driven learning [13]. It is therefore anticipated that by using these techniques on educational data, useful knowledge and information will be extracted, raising the standards of the educational system. It is important to identify learning strategies and interventions that have been proved beneficial for students to give recommendations [14].

In previous studies, it was common to launch classification and clustering approaches working with educational data. In the study of these techniques, Song et al. [9] used the support vector machine and k-means clustering, to group students based on their performance and label them as successful or not. The student record embodied 41 variables, the components of which were delinquency behaviors, academic subjects, and demographics. SVM is a process of predicting results according to clusters, while being a clustering algorithm, k-means is used to set students' socioeconomic classes and effective learning into categories.

On the other hand, Alhassan et al. [10] applied classification algorithms to their research study to explore the affinity of grades obtained in coursework and online active data in student performance. As per the research, assessment grades are the most valuable factors that can be used to predict academic performance. Besides, modeling that uses assessment grades and data on online activity also outperform others. In fact, the two algorithms that worked best were Random Forest and Decision Tree. The study emphasizes the important of incorporating instructional technologies into learning environments.

Additionally, Mohamed Nafuri et al. [11] aimed to develop clustering analysis to classify the academic performance of students in the context of Malaysian higher education and develop graduation rates. The dataset, for instance, entailed data on gender, education levels, vocational exposure, co-curricular activities, and awards. Further, techniques such as k-means, BIRCH, and DBSCAN were used when data had been prepared. Also, the best model was KMoB, which was the optimized k-mean model, and with that, five clusters of student performance were identified. Despite the fact that such stigmatization can affect the academic performance as well as the social status of individual group members, their insights showed the necessity of developing individualized educational plans with data-driven approach and predictive analytics.

In addition to the above, Queiroga et al. [12] relied on data mining techniques and evaluated data obtained from 4529 undergraduate students. The system established it by utilizing the Virtual Learning Environment, survey, and academic system data to forecast student achievement. The research revealed as highly predictive in terms of amenability to treatment, which had caused certain elements to emerge as seemingly playing a determining role, including education, subject enrollment, and neighborhood. The results were followed by specific policies of the institutions, including recruitment and reporting of resources in addition to student tracking. Besides, they also used Python with NumPy, Pandas, and Scikit-learn libraries to implement a model. This work disclosed the potential of AI to personalized learning experiences based on individual student data.

Moreover, using data mining techniques like, Naïve Bayesian, Artificial Neural Network, Support Vector Machine and Decision Tree Classifier, Li et al. [13] inspected the contributing factors to students' performance. The characteristics like attendance, gender, and nationality were detected with the help of the information of the Kalboard360 E-learning platform. The results showed that support vector machines algorithm which takes the behavioral features such as resource use and participation into consideration as the most powerful at predicting performance. The researchers explored how to integrate behavioral data in such a way as to increase accuracy of forecasting and make learning more relevant for an individual learner.

Furthermore, Ouyang et al. [14] considered the possibility to describe how AI is being used in education to navigate educational environments and to predict academic success, so that educational failure is less, and the learning process is more effective. This study, which was conducted across an online engineering course, discovered that the adoption of AI learning analytics has concrete effects like collaboration, students' satisfaction, and engagement. In fact, this paper presented the research cases that illustrated the application of AI to educational reforms, in online learning, and higher education.

Moreover, for an efficacious academic achievement and to build personal effectiveness of academics in order to possibly minimize the likelihood of academically inadequate students. Renzulli [15] has suggested several effective strategies for learning; active involvement into study content, self-testing, note-taking, time management, flexed learning, distributed practice, interleaved practice, seeking somebody, self-regulation, and using university tools and resources. This paper reveals the potential of AI technology to execute authentic applications which could assist teachers to perform well academically.

In summary, all these studies show that data mining techniques now play a crucial role in educational research through providing valuable feedback on educational data and this results in the quality improvement of the educational system. Even though such attempts have been a few, gaining access to "real-time" data and recommending learning strategies based on student performance are not yet addressed sufficiently.

III. Research Methodology

Upstream industry pioneers have produced a standard called the Cross Industry Standard Process for Data Mining (CRISP-DM). It is equally and widely utilized in a variety of domains [16]. The six stages of this standard are; Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment are its six stages [17]. In fact, its broad and thorough application in the sphere of education is particularly noteworthy because it extends beyond other domains of use [18].

In this study, a predictive model creation strategy was developed using the CRISP-DM model, providing a

comprehensive structure for project implementation. Fig. 1 shows the complete CRISP-DM steps.

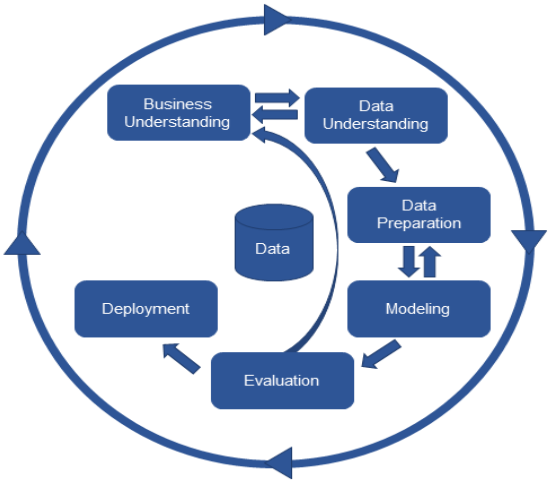

Fig. 1. Complete CRISP-DM step [17].

A. Business Understanding

This paper suggested a predictive model and recommendation system for student academic performance using educational statistics, aiming to enhance communication between educators and students, potentially enhancing learning environments [19].

As part of the methodology, attributes anticipated to have a direct impact on students' status were selected. To achieve this, a thorough investigation was conducted, followed by the construction of a figure (as shown in Fig. 2) demonstrating the relationship between these attributes and the students' attained status. Students were then classified into three categories (graduate, dropout, and enrolled) using various supervised learning techniques. Ultimately, learning tactics were suggested by the recommendation system based on the students' standing.

B. Data Understanding

The Kaggle project "Predict students' dropout and academic success" [20] provided the dataset used in this investigation. This dataset offers a comprehensive picture of the students enrolled in different undergraduate programs at a university. It contains information on social-economic characteristics, academic achievement, and demographics that can be utilized to examine potential determinants of academic success and student dropout. The dataset comprises 4424 instances, each associated with 35 attributes. The target attributes were graduate, dropout, and enrolled. The Table I shows description of the dataset including features and their type.

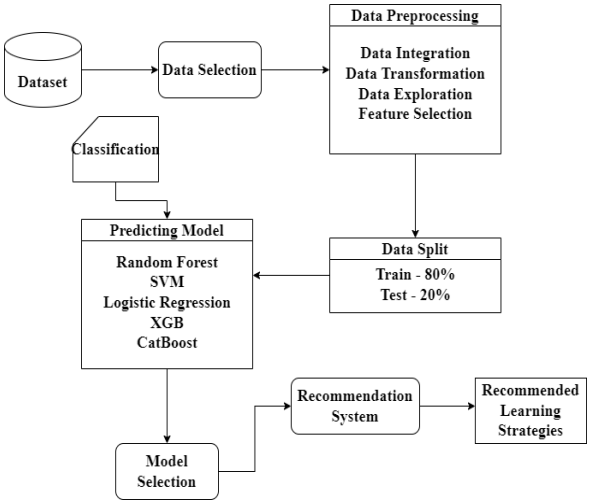

Fig. 2. Research methods for the prediction model.

TABLE I. DESCRIPTION OF THE DATASET INCLUDING FEATURES AND THEIR TYPE

Category	Feature	Description	Type
Demographic Data	Marital status	The marital status of the student	Categorical
	Gender	The gender of the student	Categorical
	Nationality	The nationality of the student	Categorical
	Age at enrollment	The age of the student at the time of enrollment.	Numerical
	International	Whether the student is an international student	Categorical
Academic Data	Course	The course taken by the student	Categorical
	Daytime/evening attendance	Whether the student attends classes during the day or in the evening	Categorical
	Curricular unit's 1st Sem (credited/ enrolled/ evaluations/approved, grade/without evaluations)	The number of curricular units credited/enrolled/evaluations/approved/ grade/ without evaluations by the student in the first semester.	Numerical
	Curricular unit's 2st Sem (credited/ enrolled/ evaluations/approved, grade/without evaluations)	The number of curricular units credited/enrolled/evaluations/ approved/grade/ without evaluations by the student in the second semester.	Numerical
	Application mode	The method of application used by the student	Categorical
	Application order	The order in which the student applied	Numerical
Social-economic Data	Mother's/Father's qualification /occupation	The qualification/occupation of the student's mother/father	Categorical
	Tuition fees up to date	Whether the student's tuition fees are up to date	Categorical
	Scholarship holder	Whether the student is a scholarship holder	Categorical
	Educational special needs	Whether the student has any special educational needs	Categorical
	Debtor	Whether the student is a debtor.	Categorical
	Unemployment rate	Unemployment rate among students	Numerical
	Inflation rate	Inflation rate of the region	Numerical
	GDP	Gross domestic product (GDP) from the region	Numerical

C. Data Preparation

Data preprocessing is a crucial step in knowledge discovery, involving cleaning, reduction, transformation, and feature selection. It involves inspecting the dataset for unwanted values, eliminating irrelevant fields, and converting the cleaned data into a suitable format for effective machine learning algorithms [16]. During this step, tasks such as checking missing data and duplicating data, replacing numerical values with names, changing categorical columns to categorical data type, getting all categorical variables except target, and rename columns were carried out.

D. Modeling

Previous studies showed that data mining involved creating models for classification, prediction, or finding hidden patterns in data that was observed [9, 10, 11]. The data mining that is prevalent these days can be divided into two types: the supervised and unsupervised data mining. However, supervised algorithms work on the basis of learning the category of unseen data, while unsupervised algorithms enable the study of the hidden patterns without target variable [19]. Logistic Regression (LR), XGBoost (XGB), Random Forest (RF), Support Vector Machines (SVM), and CatBoost (CB) classifiers were utilized as the supervised approaches to develop the models for this study.

1) Logistic regression: One prevalent linear classification approach that is effective at simulating the likelihood of a binary result of 1 or 0 is a logistic regression; it is either primarily dependent on one or more input factors. The dependent variable is a generalized combination of the predictor variables which is expressed as the logistic regression equation modeled by the logistic function, also known as the logistic or sigmoid function [21].

2) XGBoost (XGB): XGB is a supervised method which heavily relies on a specific algorithm called boosting. Starting with a learning set which includes correctly labeled samples, it is brought to a model of prediction which is used on any new examples. The training example pairs (x_0, y_0), (x_1, y_1) ... (x_n, y_n) are used by the method, where x is a vector of features and y is its label. Regression and classification issues where the label y takes a discrete or continuous value can be solved with it. The decision tree boosting technique XGBoost develops many models in a sequential fashion, each one aiming to make up for the shortcomings of the one before it. Hence, by using the XGBoost algorithm, this method can be expanded to a generalized gradient boosting [22].

3) Random Forest (RF): A Random Forest classifier is utilized for the regression and classification tasks, which involves a specific machine learning technique. It improves the accuracy by creating decision trees that are trainable on different parts of the same training set. Thus, to prevent overfitting, it learns from all the parts of the data. Random Forest generates K numbers of trees with distinct attributes each time, without pruning, by selecting attributes at random. Besides, test data is examined on every created tree, as opposed to decision trees, and the most frequent output is assigned to that particular instance [23] A forest with a larger number of trees yields the best accuracy. Missing values and category values are all handled by Random Forest. It gauges the purity and impurity of attributes using the Gini index indicator. In general, Random Forest classifiers exhibit more robustness and efficiency in comparison to decision trees [24].

4) Support Vector Machine (SVM): Vapnik's theoretical learning concept serves as the foundation for SVM. Systemic risk minimization is embodied by SVM [21]. SVMs are used in various domains related to outlier detection, regression, and classification. An SVM's original input space is mapped into a high-dimensional dot product space via a kernel. The new area is known as the feature space, and it is there that the best hyperplane for maximizing generalizability is identified. A small set of data points known as support vectors can determine the ideal hyperplane. Even though it lacks problem-domain knowledge, an SVM can produce high generalization results for classification tasks [25].

5) CatBoost(CB): A machine learning technique called CatBoost Classifier (CBC) helps in regression, classification, multi-class classification, and ranking. The objective function is reduced as the gradient decreases, and this causes variations in its design. Furthermore, built-in analytics in CatBoost evaluate the accuracy of the model before deployment. It does away with the necessity for feature processing by introducing a novel technique for managing category attributes. In addition, CatBoost makes use of ordered boosting, a permutation-driven substitute for traditional boosting techniques. Unlike the Gradient Boost, which tends to overfit quickly, it contains a mod to handle overfitting in small datasets. Additionally, CatBoost has two boosting modes: Plain and Ordered [26].

A computer with 8GB of RAM and 11th Generation Intel Core i5 processor was used for the research. Moreover, Google Colab platform was used during the entire project's development. Pandas, NumPy, Matplotlib, Seaborn, and Scikit-Learn were among the toolkits used to evaluate and compare the suggested classification models.

In addition, the dataset was manipulated in a manner that produced a 10-fold cross-validation technique which was used in splitting the dataset in training and testing subsets. Also, the stratified K-Fold subfunction for the cross-validation was employed. Scikit-Learn's model selection function [19] was used to carry out this division. The evaluations on the performance scores of classifiers were done using the cross-validation score and the GridSearchCV subfunctions.

E. Evaluation

Performance of the models are evaluated using a number of metrics, such as accuracy, precision, F1-score and recall [27]. Following metrics are used for evaluating models.

1) Precision: It is defined as the ratio of the model's total predictions to the number of true positives (TP).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

2) Recall: It is computed by dividing the total number of false negatives (FN) and true positives (TP) by the number of true positives.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

3) Accuracy: It can be calculated by dividing the total number of correctly classified examples by the total number of classified examples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

4) F1-Score: The precision and recall score averaged with respect to their weights is the F1 score. It also has consideration of false positives and fales negatives for it is to accurately indicate accuracy as well as recall.

$$\text{F1-Score} = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

TP denotes true positive: The predicted category is positive and the actual category is positive. FP designates false positive: The predicted category is positive and the actual category is negative. FN represents false negative: The predicted category is negative and the actual category is positive. TN means true negative. The predicted category is negative and the actual category is negative.

F. Deployment

The system predicted student performance such as "graduate, dropout, and enrolled" and suggested helpful strategies for learning to support them. User input features like demographic, academic data, and socio-economic data can be entered into the system.

The study utilized Python as the main programming language, Pandas for data manipulation and preprocessing, and NumPy for numerical operations. Machine learning library like Scikit-Learn was used for implementing models. Matplotlib and Seaborn were used for visualizing data distributions and performance metrics. The models were trained on an 11th Generation Intel Core i5 processor, 8GB RAM, and Google Colab notebook platform. The final recommendation system was developed and deployed using Streamlit [28], a Python-based framework for building and deploying machine learning models as web applications. The web interface allowed users to upload student data and recommend personalized learning strategies based on input.

Moreover, this research utilized some recommended learning strategies in educational psychology as follows [15].

1) Learning strategies for dropout students: The students opt to that they could dropping out should be enrolled in a soft skills program such as time management, self-testing, note-taking, and other productive study methods. Additionally, the program should be supplemented with academic counseling and follow-up to help the students consistently integrate the techniques they have learned. It is also advisable to boost active-learning approaches including creating notes and rewriting them, practicing self-testing, but not relying on other activities in passive reading. This course further should support the students in both increasing their efficiency in pursuing a reasonable time schedule and trying to devote the specified amount of time required for learning every week. Lastly, it should address any motivational problems or negative attitudes towards difficult material that may cause students to lose interest.

2) Learning strategies for enrolled students: The evaluation of students who are new to the learning system is important. The skills the students have must be evaluated and if there are any weakness, such should be identified. Once the system has provided active learning techniques, created a schedule, dealt with the obstructs, advised on the best practices and offered attention and accountability good results can be expected.

3) Learning strategies for students likely to graduate: For students who are positively engaging with learning, the system should suggest time management, note-taking, active reading, and self-testing techniques. Furthermore, it should motivate them to improve advanced skills like research, academic writing, and critical thinking.

IV. Results and Evaluation

A. Descriptive Analysis

Firstly, the influence of demographic, socioeconomic, and academic data on student performance in higher education was investigated.

1) Student performance: Enrol, Graduate, and Dropouts : The pie chart of Fig. 3 displays three segments representing student performance in higher education: Dropout (50%), Graduate (32%), and Enrolled (18%). Such a chart was used to focus on parameters of student retention and graduation rates. This may be part of a discussion with broader goals including the effectiveness of academic support systems.

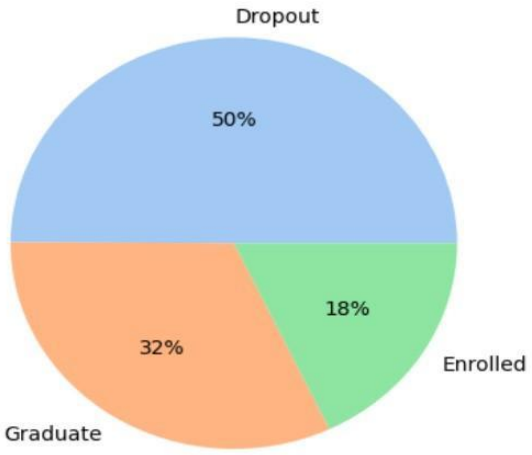

Fig. 3. Pie chart of student outcomes.

2) Top 10 factors that affect student performance: The heatmap in Fig. 4 illustrates the 10 most significant factors that affect student performance. The result of the heatmap visualized factors range from academic factors such as grades

of curricular unit and approvals to socio-economic factors such as scholarships and tuition fees.

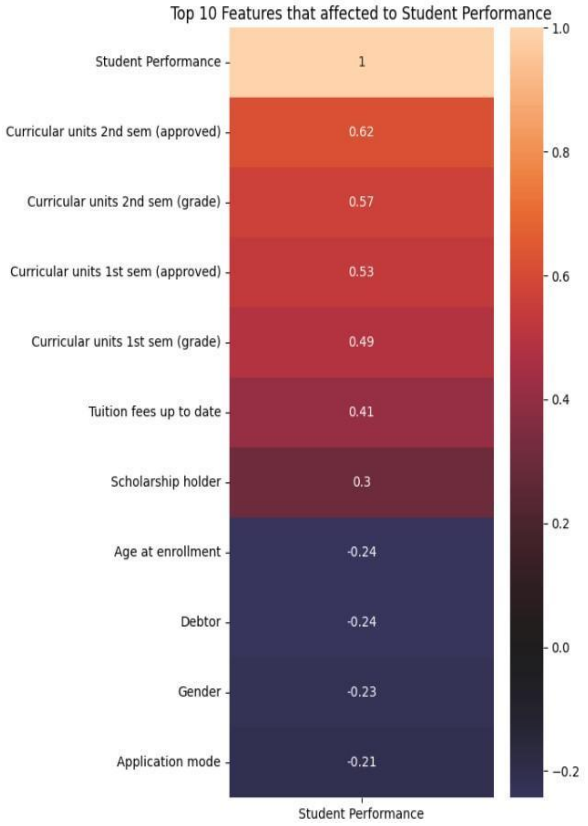

Fig. 4. Top 10 most influencing factors.

3) Age Distribution at the time of Enrollment: The histogram in Fig. 5 shows the age distribution of the students at the time of enrollment. The majority of enrolled students fell into the 18–22 age range, but it was evident that students in the 60–70 age demographic range are marginally present in the population.

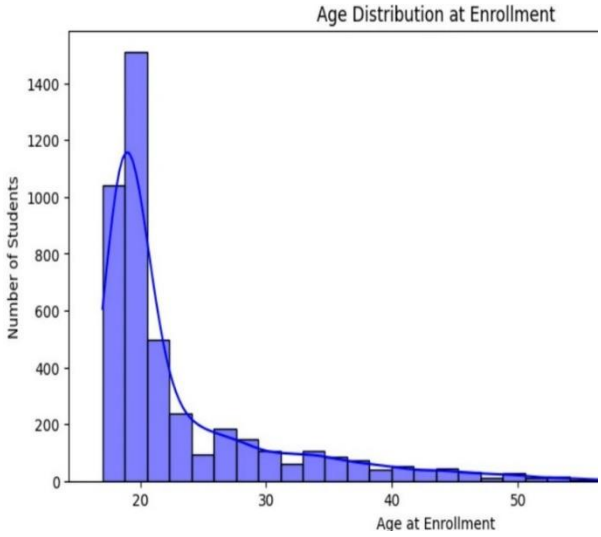

Fig. 5. Histogram of age distribution at the enrollment.

4) Effect of curricular units 2nd semester (approved) on student performance: The Fig. 6 bar chart exhibits a clear upward trend from dropouts to graduates in total curricular unit of 2nd semester (approved).

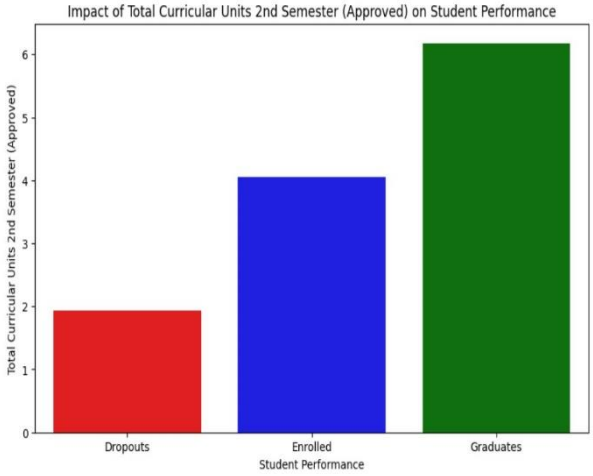

Fig. 6. Bar chart of student performance in all curricular units during the 2nd semester.

5) Effect of scholarship status on student performance: The stacked bar chart in Fig. 7 clearly illustrates a majority of scholarship holders graduating when it compared to non-scholarship holders.

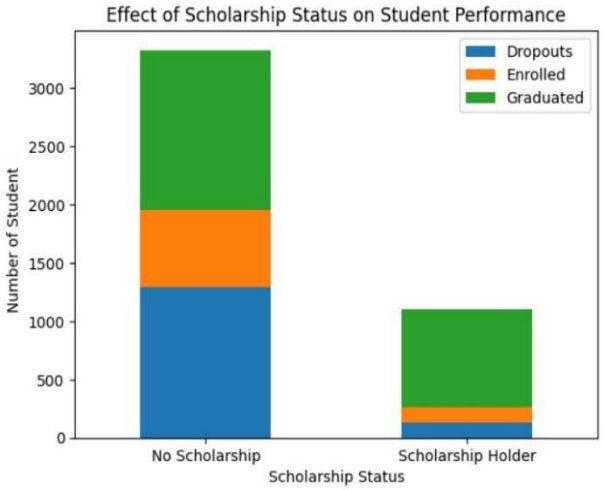

Fig. 7. Stacked bar chart of scholarship status on student performance.

B. Feature Selection

Secondly, a data mining method (classification) was applied to predict the performance of students based on affective factors. To identify the best prediction model, five distinct classifiers—Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB), CatBoost (CB), and Logistic Regression—were employed in the analysis. Each classifier was methodically assessed to determine the best outcome. Consequently, accuracy, F1-score, precision, and recall were measured, with the best results boldfaced. The results of the assessment of the selected classifiers are summarized and presented in Table II.

TABLE II. FEATURE SELECTION USING CLASSIFICATION TECHNIQUES

Algorithms	Accuracy	F1 - Score	Precision	Recall
Random Forest - RF	**0.7989**	0.7857	**0.7883**	**0.7989**
Support Vector Machine - SVM	0.7740	0.7690	0.7693	0.7740
XGBoost - XGB	0.7921	**0.7887**	**0.7883**	0.7921
CatBoost - CB	0.7955	0.7859	0.7837	0.7955
Logistic Regression - LR	0.7831	0.7686	0.7663	0.7831

With an accuracy of 79.89%, the Random Forest Classifier emerged as the leader and was now tied with the XGBoost Classifier for the highest F1-score of 78.87%. Based on these two metrics, the Random Forest Classifier was the best-performing model on the list. However, given the identical F1-scores, the XGBoost Classifier remain a formidable contender. The Random Forest Classifier is a robust and efficient method for addressing overfitting and noisy data in diverse and complex educational datasets [23]. It employs an ensemble learning approach, combining multiple decision trees, making it well-suited for handling missing values and categorical variables without extensive preprocessing, common challenges in educational data.

C. Recommendation System

Finally, a recommendation system was developed that uses a pre-trained Random Forest model to predict student performance, categorizing it into three probable statuses: enrolled, graduate, and dropout.

Based on predictions, the system delivers different learning strategy recommendations. The following are the steps in the system: Upload a csv file or manually input features, Load the pre-trained Random Forest model, predict the outcome using a model, and Display prediction probabilities and recommendations.

1) User interface: Fig. 8 depicts the web interface to get input features.

2) The Student is likely to graduate: Three categories display the prediction probabilities: 13.46%, 15.82%, and 70.72% for graduates, enrolled, and dropouts, respectively. The predicted probabilities suggested that the student had a positive chance of graduating and offered suggested learning techniques. These details are displayed in Fig. 9.

3) The Student is likely to drop out: The prediction probabilities indicate that there is a significantly larger probability dropping out at 59.80%, enrolling at 10.51%, and graduating at 29.69%. It is shown in Fig. 10.

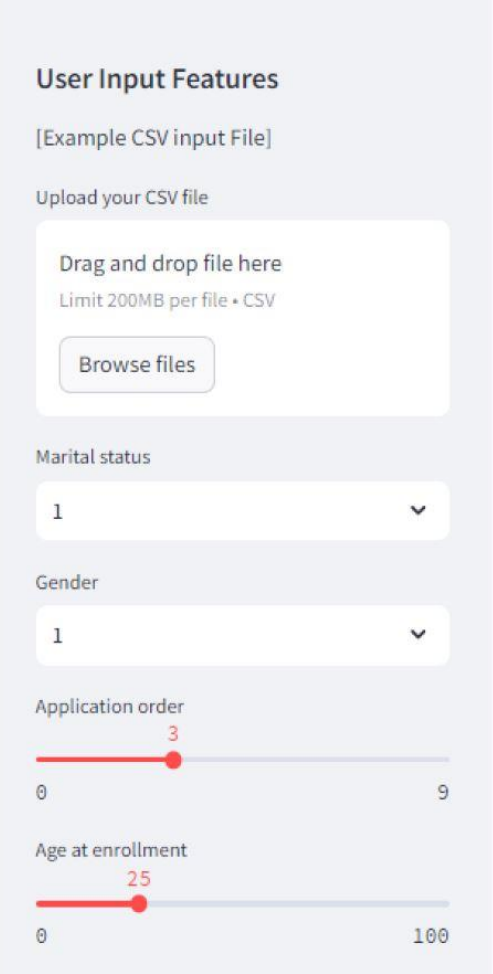

Fig. 8. Web interface to get input features.

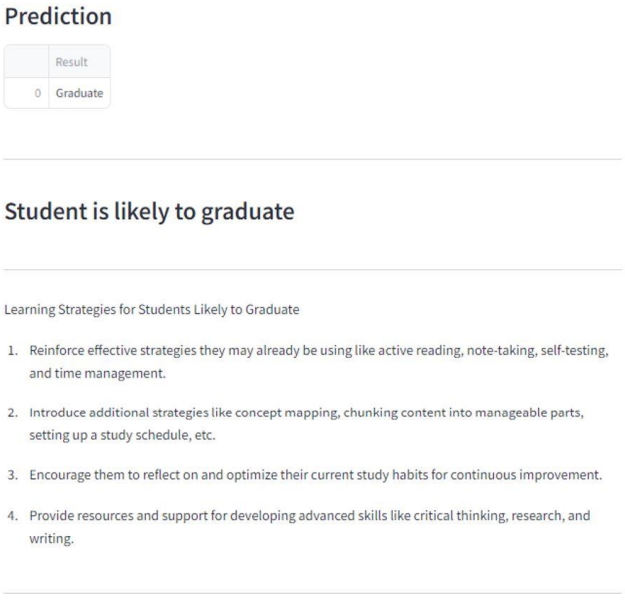

Fig. 9. Prediction probabilities and learning strategies of student is likely to graduate.

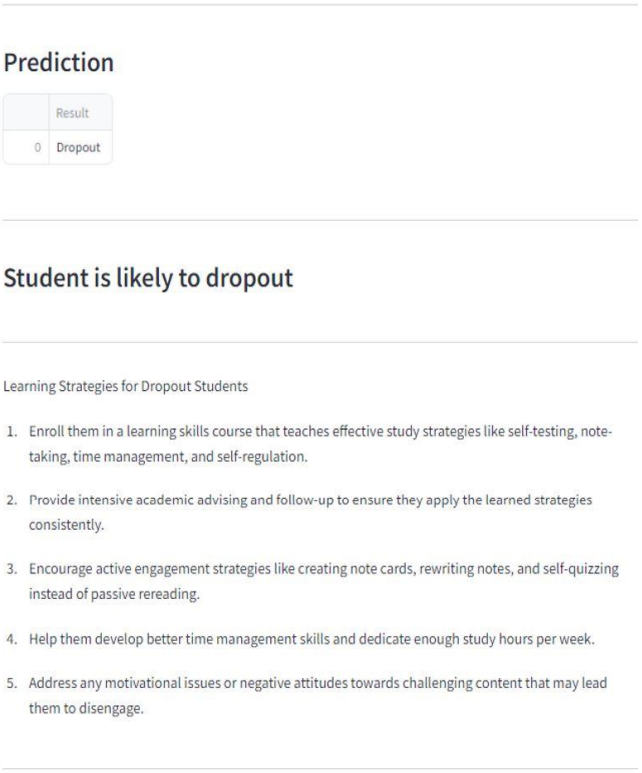

Fig. 10. Prediction probabilities and learning strategies of student is likely to dropout.

V. Discussion

In this section, the findings from the descriptive analysis, feature selection, and the recommendation system presented in the previous section are discussed.

A. The Influence of Demographic, Socioeconomic, and Academic Factors

This article generally indicates the vital part that demographic, socioeconomic and academic factors play in the tertiary education performance of the students. Contrary to the graduates (32%) and those who are still enrolled (18%), the very high dropout rate (50%) demonstrated that higher education system has a serious issue. This distribution in fact suggests that outside factors that might be related to socioeconomic status and demographic backgrounds, play a significant role in the outcomes of learners even when there are educational programs.

The factors that affect student performance further support the complex relationship between student success and socio-economic factors, such as scholarships, tuition costs, and academic indicators like the number of curricular units in semesters. This implies a need for assistance with both academic challenges and economic hurdles. Consequently, the results of this research are also supported by the findings of [9, 13]. In Song et al. [9], the authors explored issues including behavior, academics, and demographics. Behavioral components englobed being absent from class, while academic achievement, coupled with good grades was fundamentally important. Socioeconomic status and year of entry were the demographic elements involved. Even during Li et al. [13], the authors went for the details of demographic variants like age, occupations of parents, and nationality, which connected the rich diversity of students with the advantages and challenges of different cultures. The socioeconomic factors, like parents' level of education or income of the family, influenced their education, the availability of resources and their academic success to a great extent. Academic factors, one of which is previous performance and absences, are directly correlated with the student performance and therefore, they have a chance to predict future outcomes.

B. Efficiency of Classification Methods in Predicting Student Performance

Applications of many classifiers, the obtaining of insightful results, and the prediction of student achievement have been done. The Random Forest classifier [23] has an accuracy of approximately 80%; it performs better than other models, and what makes it robust is its ability to efficiently handle many different kinds of datasets that are common in educational settings. Random Forest's performance is an indication that such ensemble learning approaches are essential for successful educational data mining. These results are invaluable for developing algorithms that can identify at-risk student cohorts, enabling timely interventions.

The study [10] employs different types of classification for the prediction of student success and the grade of assessment is revealed to be the major factor in rates of performance. The models that were using the grades and online activity data performed better. However, it was the Random Forest classification model that performed better in predicting student performance.

C. Implications of Recommendation System

The development of a recommendation system that classifies students to likely categories (graduate, enrolled, or dropout) and provides specific educational strategies based on the predictions is especially inventive. This technology is very promising, it puts machine learning and its predictive power to real use in a way that has a practical impact on the education process, and can completely change the concept of teaching and learning. E.g., 59.80% dropout rate of this student could trigger the implementation of special supportive measures like tutoring, counseling, or modification of financial aids by the academic institutions.

In addition, a similar conclusion was reached in studies [14, 15], which emphasize the significance of using learning analytics with AI-based performance prediction models. The results of such a technique demonstrate that cooperation is enhanced, that satisfaction is higher and that engagement and learning are improved. Such study shows the trend toward educational effectiveness by emphasizing the application of AI models which have been created and the gap bridging between the creation and use of AI models.

D. Implications and Limitations

By personalizing progressions of study to suit the needs of individual students, offering predictive insights for early intervention and continuously enhancing instructional methodologies and techniques, AI can improve personalized

learning [15]. It could thus lead to improved academic work, a drop in cases of dropout rate as well as more effective learning.

The amount, class, and constant quality of data, which are available, determine how efficient AI systems work [29]. They might tarnish the trust because the terms are hard to be understood or difficult to comprehend. Differing from one to the other in practices and demography could limit the applicability of the same interventions in various settings [14]. In fact, information security and privacy are among the dilemmas that ethical corporations often encounter [30]. If AI extensively used, task automation will supplant people interaction and people will experience threats of a decline in critical thinking and social skills. The issue of resistance, a lack of technical compatibility, and the need for educating the staff may complicate automated systems integration with existing educational infrastructure.

VI. Conclusion

This study titled "EDM and AI-Powered Learning Strategy Recommendation System to Boost the Academic Results" articulates that AI and EDM can be used in learning settings to improve academic performance. Students' results prediction model and recommendation engine which use variables like socioeconomic status, demographic data source and academic performance was built using the CRISP-DM model. Notably, in this case, this research reconciled this navigation with machine learning algorithms in a perfect way by proving that random forest was the ideal classifier for the student's performance status prediction. With the implied procedure proposed as a basis, unique study systems were built for the learners on the foundation of their particular needs, desires, and struggles. The fact that personalized education might be achieved within these systems is a crucial milestone for the field of personalized learning in general. The evidence has a significant impact that will aid institutes to formulate student progress and knowledge exploration concerning factors that invite student success. In this regard, the AI-powered system which would sense learners and respond with personalized instructions should promote the students' access rights, involvement, and productivity. Through this shift, the numbers of student dropout might come down and a comfortable atmosphere in which the learning process can be done effectively and this could be realized.

In the future, real-time data analytics integration and ongoing AI algorithm improvement should improve on assessing the productivity and adaptability of learning systems. Moreover, investigating hybrid models that fuse AI suggestions with the knowledge of human educators may result in more comprehensive teaching approaches.

References

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art ," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601-618, Nov. 2010, doi: 10.1109/TSMCC.2010

[2] C. Romero and S. Ventura, "Data mining in education," WIREs Data Mining and Knowledge Discovery, vol. 3, pp. 12-27, 2012, doi: 10.1002/widm.1075

[3] A. Dutt, M. A. Ismail and T. Herawan, "A Systematic Review on Educational Data Mining," in *IEEE Access*, vol. 5, pp. 15991-16005, 2017, doi: 10.1109/ACCESS.2017.2654247

[4] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," Journal of Medical Systems, vol. 43, no. 6, Apr. 2019. doi:10.1007/s10916-019-1295-4.

[5] D. Hooshyar, M. Pedaste, and Y. Yang, "Mining educational data to predict students' performance through procrastination behavior," *Entropy*, vol. 22, no. 1, p. 12, Dec. 2019. doi:10.3390/e22010012

[6] T. Liu, C. Wang, L. Chang, and T. Gu, "Predicting high-risk students using learning behavior," *Mathematics*, vol. 10, no. 14, p. 2483, Jul. 2022. doi:10.3390/math10142483

[7] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, "Identification of student behavioral patterns in higher education using K-means clustering and support vector machine," *Applied Sciences*, vol. 13, no. 5, p. 3267, Mar. 2023. doi:10.3390/app13053267

[8] T. Tao, C. Sun, Z. Wu, J. Yang, and J. Wang, "Deep neural network-based prediction and early warning of student grades and recommendations for similar learning approaches," *Applied Sciences*, vol. 12, no. 15, p. 7733, Aug. 2022. doi:10.3390/app12157733

[9] Z. Song, S.-H. Sung, D.-M. Park, and B.-K. Park, "All-Year Dropout Prediction Modeling and Analysis for University Students," Applied Sciences, vol. 13, no. 2, p. 1143, Jan. 2023, doi: 10.3390/app13021143.

[10] A. Alhassan, B. Zafar, and A. Mueen, "Predict students' academic performance based on their assessment grades and online activity data," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020. doi:10.14569/ijacsa.2020.0110425

[11] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," Applied Sciences, vol. 12, no. 19, p. 9467, Sep. 2022, doi:10.3390/app12199467.

[12] E. M. Queiroga et al., "Using Virtual Learning Environment Data for the Development of Institutional Educational Policies," Applied Sciences, vol. 11, no. 15, p. 6811, Jul. 2021, doi: 10.3390/app11156811.

[13] F. Li, Y. Zhang, M. Chen, and K. Gao, "Which Factors Have the Greatest Impact on Student's Performance," Journal of Physics: Conference Series, vol. 1288, p. 012077, Aug. 2019, doi: 10.1088/1742-6596/1288/1/012077.

[14] F. Ouyang, M. Wu, L. Zheng, L. Zhang, and P. Jiao, "Integration of Artificial Intelligence Performance Prediction and learning analytics to improve student learning in online engineering course," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Jan. 2023. doi:10.1186/s41239-022-00372-4

[15] S. J. Renzulli, "Using learning strategies to improve the academic performance of university students on academic probation," NACADA Journal, vol. 35, no. 1, pp. 29–41, Jul. 2015. doi:10.12930/nacada-13-043

[16] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," J. Data Warehousing., vol. 5, no. 4, pp. 13–22, 2000.

[17] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISPDM process model," Procedia Comput. Sci., vol. 181, pp. 526–534, 2021. doi:10.1016/j.procs.2021.01.199

[18] W. F. Yaacob, S. A. Nasir, W. F. Yaacob, and N. M. Sobri, "Supervised Data Mining Approach for predicting student performance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, p. 1584, Dec. 2019. doi:10.11591/ijeecs.v16.i3. pp1584-1592

[19] M. Bellaj, A. Ben Dahmane, S. Boudra, and M. Lamarti Sefian, "Educational Data Mining: Employing Machine Learning techniques and hyperparameter optimization to improve students' academic performance," International Journal of Online and Biomedical Engineering (iJOE), vol. 20, no. 03, pp. 55–74, Feb. 2024. doi:10.3991/ijoe.v20i03.46287

[20] "Predict students' dropout and academic success," Kaggle, https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention (accessed Apr. 18, 2024).

[21] A. Salah Hashim, W. Akeel Awadh, and A. Khalaf Hamoud, "Student performance prediction model based on supervised machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, p. 032019, Nov. 2020. doi:10.1088/1757-899x/928/3/032019

[22] R. Mitchell and E. Frank, "Accelerating the XGBOOST algorithm using GPU computing," *PeerJ Computer Science*, vol. 3, Jul. 2017. doi:10.7717/peerj-cs.127.

[23] N. S. Ahmed and M. Hikmat Sadiq, "Clarify of the random forest algorithm in an educational field," *2018 International Conference on Advanced Science and Engineering (ICOASE)*, Oct. 2018. doi:10.1109/icoase.2018.8548804

[24] M. M. Tamada, R. Giusti, and J. F. Netto, "Predicting students at risk of dropout in technical course using LMS Logs," *Electronics*, vol. 11, no. 3, p. 468, Feb. 2022. doi:10.3390/electronics11030468

[25] Y. Chen, "Support Vector Machines and Fuzzy Systems," *Soft Computing for Knowledge Discovery and Data Mining*, pp. 205–223, 2008. doi:10.1007/978-0-387-69935-6_9

[26] A. Joshi *et al*., "CatBoost — an ensemble machine learning model for prediction and classification of student academic performance," *Advances in Data Science and Adaptive Analysis*, vol. 13, no. 03n04, Jul. 2021. doi:10.1142/s2424922x21410023

[27] G. Feng, M. Fan, and C. Ao, "Exploration and visualization of learning behavior patterns from the perspective of educational process mining," *IEEE Access*, vol. 10, pp. 65271–65283, 2022. doi:10.1109/access.2022.3184111

[28] M. Khorasani, M. Abdou, and J. Hernández Fernández, "Streamlit Basics," *Web Application Development with Streamlit*, pp. 31–62, 2022. doi:10.1007/978-1-4842-8111-6_2

[29] M. Murtaza, Y. Ahmed, J. A. Shamsi, F. Sherwani, and M. Usman, "AI-based personalized e-learning systems: Issues, challenges, and solutions," IEEE Access, vol. 10, pp. 81323–81342, 2022. doi:10.1109/access.2022.3193938

[30] K. Zhang and A. B. Aslan, "AI Technologies for Education: Recent research & future directions," Computers and Education: Artificial Intelligence, vol. 2, p. 100025, 2021. doi:10.1016/j.caeai.2021.100025