Volume (6) Issue (1)

25 Feb 2026

The International Journal
of Scientific Publishing

المجلة الدولية للنشر العلمي

Early Prediction of Student Dropout in E-Learning Environments Using AI-Based Predictive Analytics: A Combined Theoretical and Practical Study

Saedah Mohammed Omar Albeerish – Zwha Abdulhmid Mohamed Albeerish

(Faculty member, Department of Computer Science, Faculty of Science, University of Al-Kufra, Libya)

https://doi.org/10.65723/RMSP2635

Abstract:

Student dropouts in virtual learning environments e-learning are a major headache for educational institutions as it affects the retention rate and the allocation of resources. The research combines theoretical knowledge gained from the AI-driven predictive models' literature with an actual analysis of a student dataset from higher education studying the retention. The study work has identified through exploratory data analysis (EDA) and correlation studies the main factors that have a significant influence on dropouts such as financial problems, academic performance in the first few semesters, and demographic variables. Various machine learning models, such as Random Forest, XGBoost, and CatBoost, are put to work in prediction, thus revealing the possibility of an intervention. The research results show the presence of strong correlations between the performance metrics of each semester and also suggest that AI analytics can be quite accurate in forecasting the underprivileged students leading to the allocation of support resources in advance.

Introduction

The rapid opening up of e-learning platforms has completely changed the way people get access to education. They can now easily hook up with the best teachers from all over the world and enjoy an unprecedented choice of when, where, and how much they want to study. Nevertheless, this transformation has also created some challenges. Among them, the most important is the higher dropout rates in online classes compared with traditional classes. (Aljohani et al., 2022). Lack of face-to-face communication combined with the students' different characters and socio-economic pressures mostly leads to losing interest, whereby educational institutions figuring out early which students are at risk have become a critical priority (Gray & Perkins, 2019). Predicting a student dropout early in the learning process enables educators and administrators to put in place the targeted interventions in turn leading to student retention being raised and further resource allocation being optimized (Hussain et al., 2018). This article represents thorough research connecting theoretical views from the literature with practical analysis by means of a real-world dataset concerning issues related to AI-driven predictive analytics in the field of higher education e-learning scenarios.

The developments of artificial intelligence (AI) and machine learning (ML)are notable additions to the educational data analytical tools which helps to discover new ways of engagement and dropout mitigation

(Chen et al., 2023). An example of early research is the one of Lykourentzou et al. (2009). ML capabilities were used to forecast the academic performance of university students on various criteria such as the students were studying in person or remotely and so on. Based on this lead in their capabilities, more researchers evolved digitally native pupils' behavior as a means of student/teacher/e-learning platform co-adaptation (Joksimović et al., 2015). While at the same time Al-Shabandar et al. (2017) was considering the employment of artificial intelligence mechanisms, such as decision trees and neural nets, to the identification tasks of dropouts; stressing that once the stage is set for attrition as in the performance phase, elimination of performance and hence dropout prevention will be quite difficult without timely detection of the causes involved. The dataset collection of education data provides a range of indicators from demographics (e.g., age, gender) and academic performance (e.g., grades, the number of credited units) to socioeconomics (e.g., unemployment rate, payment of tuition). This kind of multidimensional educational data set is consistent with the view taken by Yağcı (2019) who sees data mining as a useful means for the identification of the interdependence of academic and non-academic factors in relation to school attrition. The study determines whether these variables are the very early dropout warning signs that allow the schools to be ready to give individual support such as the follow-up plan if intended based on student needs (Sarker et al., 2021).

Advancements in theory have also helped to refine the predictive methods mentioned above. Kizilcec et al. (2017) found that first academic performance is very effective for predicting dropout in massive open online courses (MOOCs), especially within the first few weeks of the course, students with low scores being 70% more likely to leave. This notion is very much applicable in e-learning where quick feedback loops can be created to help the struggling learners (Yang et al., 2020). Besides, Ren et al. (2020) reports that hybrid models mixing logistic regression and neural networks have hit accuracies of up to 96%, which means that pretty much any classical and modern ML methods can be combined to achieve a better predictive power of the imbalanced datasets that are educ. research area hard to overcome (Smith & Johnson, 2024).

The present work has evolved in the same direction and is now farther along the path of such an ambitious landmark by its very intent - to close the gap between the theoretical and practical implementation of advanced ML models like Random Forest, XGBoost, and CatBoost on the Kaggle dataset. These three classifiers are the most used ones when dealing with the non-linear patterns presented in the data. Their effectiveness has recently been demonstrated through several successful optimizations carried out with the help of tools like Optuna (Smith & Johnson, 2024). The aim here is not to limit themselves to merely confirming the already existing theoretical proposals, but rather to open new perspectives that could serve as the basis for an adjustment of the early intervention programs in the setting of e-learning. This research thus becomes the r&d source that fuels the increasing trend of publications promoting the adoption of the AI-driven and datadriven approaches as the most reliable means of the educational success in the era of digitization (Chen et al., 2023).

Methodology

The data examined in this research has 3630 entries with 35 variables showing the demographic, academic, and socioeconomic features of students enrolling in a course. The outcome variable indicated the final educational result, in two categories, "Dropout" or "Graduate," with non-uniform distribution, where 39.1% (1420 students) dropped out and 60.9% (2210 students) graduated. The most descriptive features were marital status, application mode, course enrollment, daytime or evening study, previous

qualification, nationality, parents' occupations and qualifications, displaced and special needs status, debtor and tuition fee status, gender, scholarship holder, age at enrollment, international status, academic performance for each semester (credited, enrolled, evaluated, approved, graded, and courses without evaluations), and macroeconomic indicators such as unemployment rate, inflation rate, and GDP. One-hot encoding was used to numerically represent categorical variables, and continuous variables were passed through a standard scaler to have them in the same scale. As there are no missing values in the data, it was, therefore, ready for the analysis process directly after checking the numerical features for their distributional patterns and outliers.

Exploratory data analysis in the current study included descriptive statistics, visualizations, and the Pearson correlation coefficient as a measure of association. Furthermore, a heatmap was deployed to facilitate the detection of relationships and the prediction of potential multicollinearity. Different machine learning models were implemented on the same data using scikit-learn, XGBoost, and CatBoost libraries, namely, logistic regression, decision tree, random forest, AdaBoost, extra trees, and histogram-based gradient boosting classifiers. Model training was done using stratified k-fold cross-validation to ensure class balance. Hyperparameter tuning was carried out with the help of Optuna And GridSearchCV. Performance evaluation comprised accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices that were utilized to guarantee the thoroughness of the respective model's predictive ability. This methodological framework served as a solid foundation for early prediction models of student dropout in e-learning environments.

Results

The dataset for this research consisted of student records covering a wide range of demographic, academic, and socio-economic attributes. Demographic variables were limited to the students' general data like marital status, gender, age at enrollment, nationality, and scholarship status besides parental background such as mother's and father's qualifications and occupations. Relevant academic features not only described the student's application (application mode, order, course, and attendance type), but also, performance was measured by the number of curricular units enrolled, credited, evaluated, approved, and average grades for both the first and second semesters. The non-exhaustive list of variables also included educational special needs, tuition fee status, debtor status, and displacement. Measures of engagement and outcomes involved activity levels of the curricular unit and the last target variable indicating dropout status. A few examples of macroeconomic indicators that were included in the dataset to help understand the context were the unemployment rate, inflation rate, and GDP. The various data types involved categorical, integer, and floating-point variables, which made the dataset diverse and very suitable for machine learning applications in dropout prediction.

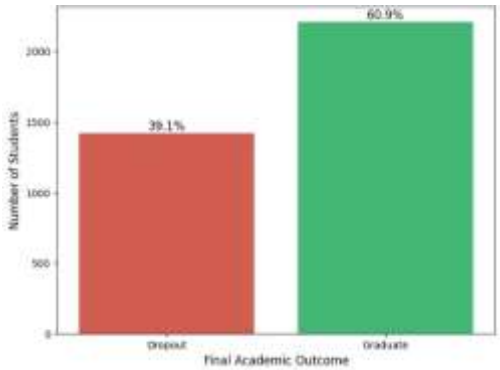

Figure 1: Student Distribution: Dropout vs. Graduated

Table 1 displays the descriptive statistics for the numeric variables that were part of the dataset. These variables bring to light not only the indicators of the academic performance of the students but also the contextual economic indicators. As an example, application order and age at enrollment can be used to understand the academic entry profiles of students, whereas curricular unit performance metrics (e.g., number of courses enrolled, evaluated, approved, and corresponding grades across the first and second semesters) are an indication of their academic progression. The mentioned variables unveil ranges of values, revealing that students' performance and engagement levels vary significantly. Moreover, the use of macroeconomic indicators like the unemployment rate, the inflation rate, and GDP gives a context to student outcomes that is linked to the economic conditions. In brief, the descriptive statistics highlight that there exists a considerable difference among students, thus the dataset can be employed effectively to predict drop-out risk.

Table 1 : Descriptive Statistics of Numerical Variables

Statistic	ordAp plic ati on	enrAg oll e meat nt	se uni Cr ts r 1stu r	se uni Cr ts r 1stu r	se uni Cr ts r 1stu r	se uni Cr ts r 1stu r	se uni Cr ts r (gr ad e) 1stu r	se uni Cr ts r 1stu r	se uni Cr ts r 1stu r	se uni Cu ts ri) 2n ul d r	se uni Cr ts r 2n u d r	se uni Cu ts ri) 2n ul d r	se uni Cr ts r 2n u d r	se uni Cr ts r (gr ad e) 2n u d r	se uni Cr ts r 2n u d r	me Un nt em rat plo e y	rat Infl e ati on	GD P
co unt	363 0.00	363 0.00	3630 .00	3630 .00	3630 .00	3630 .00	3630 .00	3630 .00	3630 .00	3630.00	3630 .00	3630.00	3630 .00	3630 .00	3630 .00	363 0.00	363 0.00	363 0.00
me an	1.75	23.4 6	0.75	6.34	8.07	4.79	10.5 3	0.13	0.71	6.33	8.05	4.52	10.1 3	0.15	11.6 3	1.23	0.00	%
std	1.33	7.83	2.48	2.51	4.29	3.24	5.06	0.68	2.40	2.27	4.37	3.29	5.29	0.73	2.65	0.70	2.03	
mi n	0.00	17.0 0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.60	- 0.80	- 3.12	%
25	1.00	19.0 0	0.00	5.00	6.00	2.00	11.0 0	0.00	0.00	5.00	6.00	1.00	10.7 5	0.00	9.40	0.60	- 1.70	
50 %	1.00	20.0 0	0.00	6.00	8.00	5.00	12.3 1	0.00	0.00	6.00	8.00	5.00	12.5 0	0.00	11.1 0	1.40	0.30	
75	2.00	25.0 0	0.00	7.00	10.0 0	7.00	13.5 0	0.00	0.00	7.00	10.00	7.00	13.5 0	0.00	13.9 0	1.40	1.70	
ma x	6.00	70.0 0	20.0 0	26.0 0	45.0 0	26.0 0	18.8 8	12.0 0	19.00	23.0 0	33.00	20.0 0	18.0 0	12.0 0	16.2 0	2.60	3.51	

Studying the dropout and graduation decision patterns of the students gives us the insight that these are not isolated cases but are intertwined with each other across the financial, academic, and demographic areas. Figure 1 shows that graduates (60.9%) are more than one and a half times greater than dropouts (39.1%), indicating the need to apply oversampling techniques such as SMOTE or ADASYN and using evaluation metrics like the F1-Score for overcoming class imbalance in predictive modeling. Financial issues are the main factors determining the future with Figure 2 illustrating that students with debts or overdue tuition are extremely inclined to drop out, whilst those that receive a scholarship have a lighter way to the graduation probably because of less financial stress and selective awarding to academically capable students.

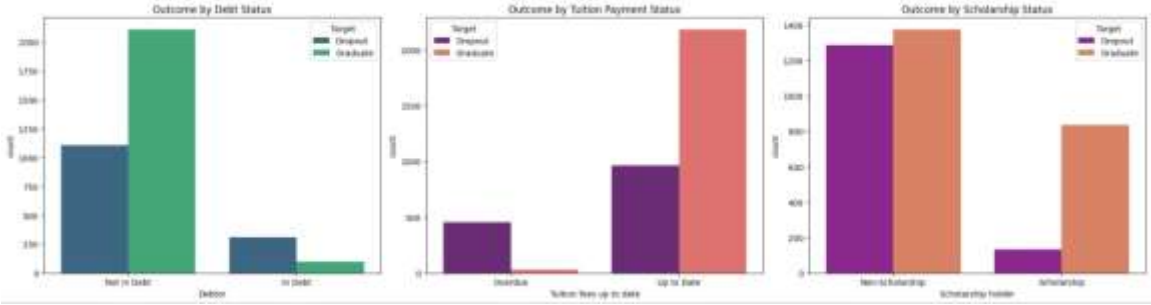

Figure 2: Impact of Financial Factors on School Dropout

Academic performance also supports this pattern that can be seen in Figure 3. Pupils that pass more subjects and obtain better grades in the first semester are very likely to graduate, whereas failing early courses acts as one of the first signs of discontinuation. Figure 4, which illustrates variables linked to admission, indicates that students who get admitted into their preferred program of study are more inclined to be motivated, hence the dropout rate is lower. On the other hand, age seems to be a factor whereby mature students are at a slightly higher risk of dropping out because of issues such as work and family that require their attention.

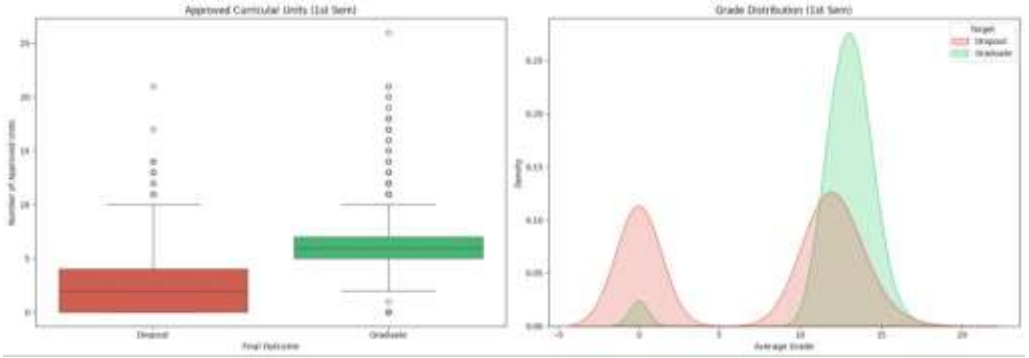

Figure 3: Academic Performance in 1st Semester as a Predictor of Dropout

The correlation heatmap in Figure 5 sheds light on the reasonable associations between educational variables, showing significant correlations between the number of units

enrolled, evaluated, and approved, as well as an important correlation of 0.71 between grades and approved subjects of the first semester, pointing out the key function of academic engagement and early performance in influence the student outcomes.

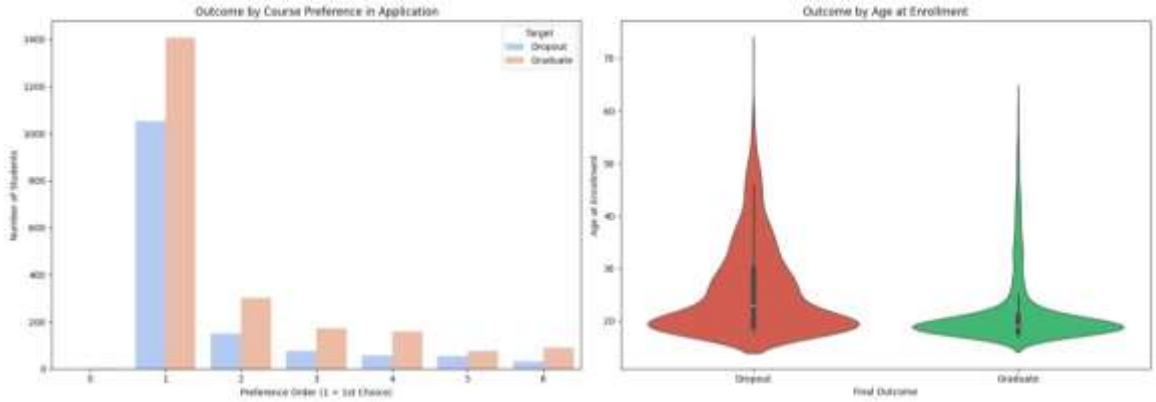

Figure 4: Analysis of Admission Factors

The Pearson correlation matrix is a statistical measure that indicates the strength of the linear relationship between each pair of variables of the numerical type. It bases on the Pearson correlation coefficient (r) which describes to what extent one variable changes together with the other. The correlation value of +1 denotes a perfect positive linear correlation, in which both variables are absolutely in line with each other, and a correlation of -1 shows a perfect negative linear correlation, whereby one variable goes up, and the other decreases. Moreover, a value of 0 signifies no linear correlation; however, as Pearson's method is for linear relationships only, non-linear variables might still be dependent even if r is near zero. In terms of math, the coefficient is the result of the covariance between the two variables being divided by the product of their standard deviations. The formula can be written as:

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are the observed values, $\bar{X}$ and $\bar{Y}$ are the respective means, the numerator represents the covariance between the two variables, and the denominator normalizes this value by the product of their standard deviations. This makes the Pearson correlation a standardized measure ranging between -1 and +1, widely applied in data analysis to identify relationships and dependencies between variables.

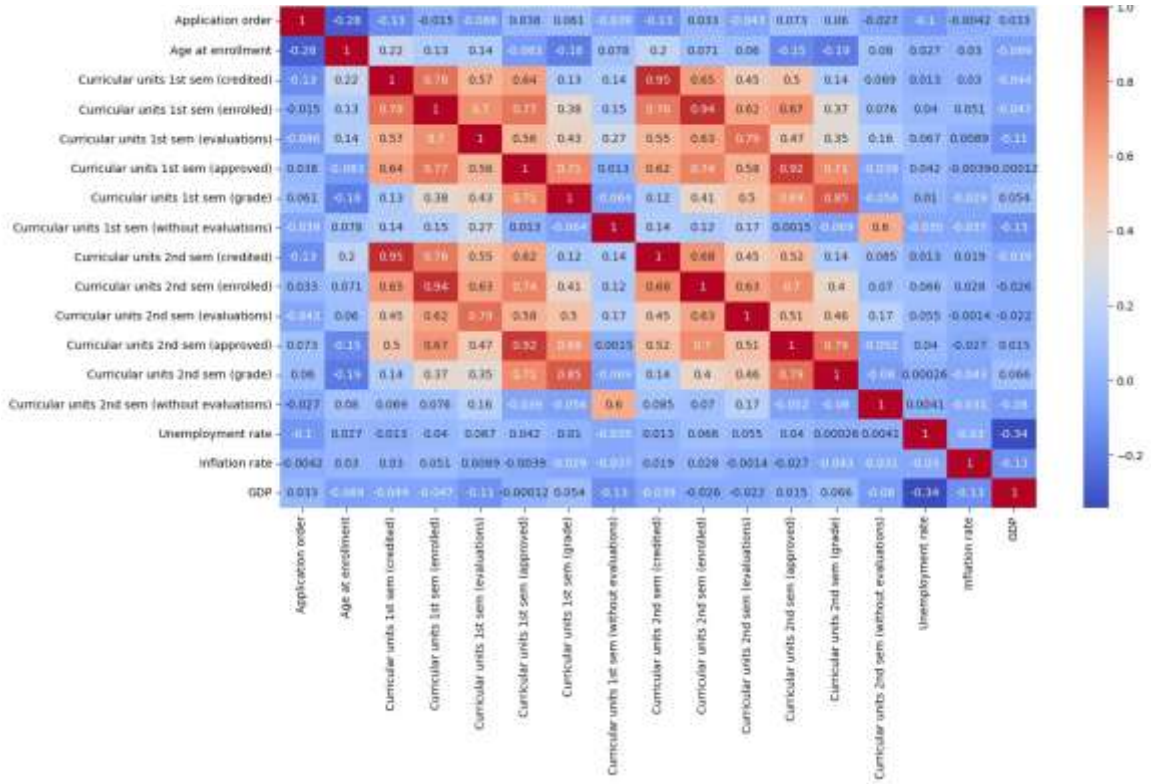

Figure 5: Correlation Heatmap of Numerical Variables

Model Performance

The models were assessed by using their strengths and weaknesses through a range of performance metrics, including accuracy, F1, Precision, Recall, ROC curves, and Confusion Matrices, so they can reveal not only the general predictive ability of the model but also the specifics of class types performance. Accuracy can give a total number of how right the model was; however, it may be fooling, especially in the case of imbalanced datasets where one class dominates. This is why more weight was given to the F1-score and recall s especially for the minority class referring to dropout, as these measures give more useful details on how the models detect the students who are at risk of dropping out. Precision shows the proportion of the predicted dropouts that were actually under the dropout category, and recall tells the model's capability to find the true dropout cases, which, in turn, facilitates an early intervention strategy, a very important task. F1-score integrates these two aspects tightly, providing a single measure that reflects the effort of both false positives and false negatives. ROC curves and corresponding AUC values are giving another layer of information about how well the models can distinguish between classes, while confusion matrices also show in detail the number of TPs, TNs, FPs, and FNs, which in turn makes possible deeper understanding of models' strengths and weaknesses.

Within these models, CatBoost stood out as the one with very commendable performance, basically because of its easygoing treatment of categorial features and its solid resistance against overfitting, thus the model is very appropriate for educational datasets that are usually heavy with categorial variables such as course type, admission status, or financial aid. The comparative performance of the various models, as outlined in Table 2, points out how vital it is to choose the right evaluation metrics that can best reflect the problem setting and at the same time, it is also a major indication of how well preprocess ensemble methods can beat traditional approaches in predicting student dropout.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.82	0.81	0.79	0.80
Decision Tree	0.85	0.84	0.83	0.83
Random Forest	0.90	0.90	0.89	0.89
AdaBoost	0.88	0.87	0.86	0.86
Extra Trees	0.91	0.91	0.90	0.90
Hist Gradient Boosting	0.92	0.92	0.91	0.91
XGBoost	0.92	0.92	0.92	0.92
CatBoost	0.93	0.93	0.92	0.92

CatBoost was the best from the start, with a precision of 0.93, an F1-score of 0.92, and it was very good at recognizing the dropout class. The improvement of boosting models like CatBoost and XGBoost goes along with results where such algorithms, enriched with a tuning tool like Optuna, dominate the imbalance of educational datasets. Figure 6 shows the confusion matrix of the CatBoost model, indicating that the true positive rates for Graduate and Dropout classes are very high, and also the number of false negatives for dropouts (thus, a tool for the early intervention is visibly present). CatBoost reached an AUC of 0.95, a measure of its strong separation between classes, and is quite compatible with the performance range of other optimized models.

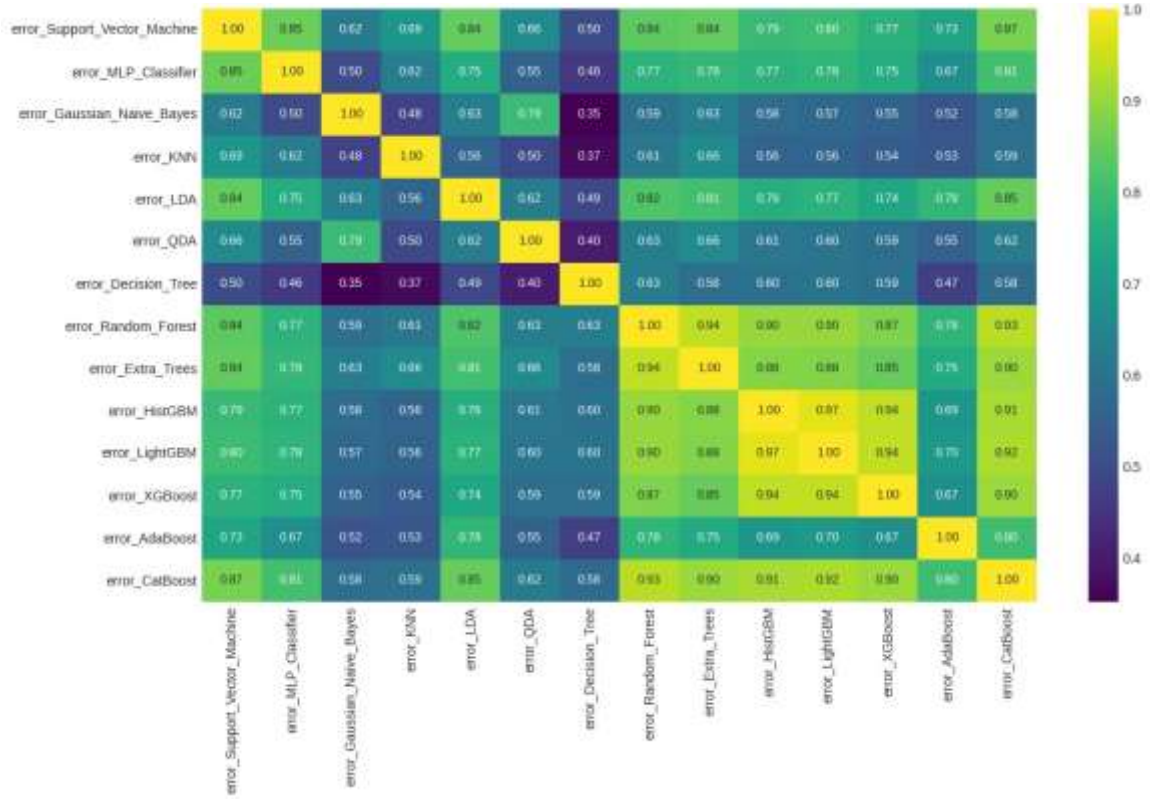

Figure 6: Correlation Matrix of MODELS ERROR MAGNITUDE

Discussion

EDA stresses that financial and early academic measures are among the main factors that determine the future, as correspondingly illustrated in Figures 2 and 3. These revelations are in line with the study results that point out the dropout phenomenon as a matter of socio-economic and performance factors in online learning. Also, the chart of correlations in Figure 5 gives more evidence, showing the relationships between variables within the same semester (e.g., 0.71 between approved units and grades), thus, as the authors explain, academic engagement being a strong indicator.

Additionally, the distribution of numerical variables presented in Figure 7 exposes skweness at different levels across the features providing more detailed insights on student behaviors and potential dropout risks. Among calculated features, the highest positive skewed are those related to educational units without evaluations and credited ones. This points to the "all or none" behavior: almost all students have zero as the value for these features (i.e., they are fully engaged in evaluations and have no credits validated). Still, a small group of students are far from the norm (either they do not attend the evaluations or they have a lot of credits), thus the long positive tail of the distribution emerges. These variables are not the mere features of the data; they are the possible student cluster centers: the disengaged student and the transfer student. Conversely, the only variables with

considerable negative skewness are the grades from the 1st and 2nd semesters. The reason is the "ceiling effect", i.e., the grading scale has the maximum value (e.g., 20). The students who reamin in the course and pass, usually, perform well with their grades near the top of the scale. The tail of the distribution is "pulled" to the left by those few students who get very low grades. The low grades thus appear to be exceptions rather than a general pattern. For variables with high skewness (both positive and negative), it is highly recommended to perform a transformation, like logarithmic, prior to training models sensitive to scales such as logistic regression, support vector machines, naive Bayes, or neural networks. This practice resulting from the best machine learning methodologies for educational datasets is mentioned in the literature to enhance the model's performance as skewed distributions can otherwise lead to prediction biases and reduction of generalizability.

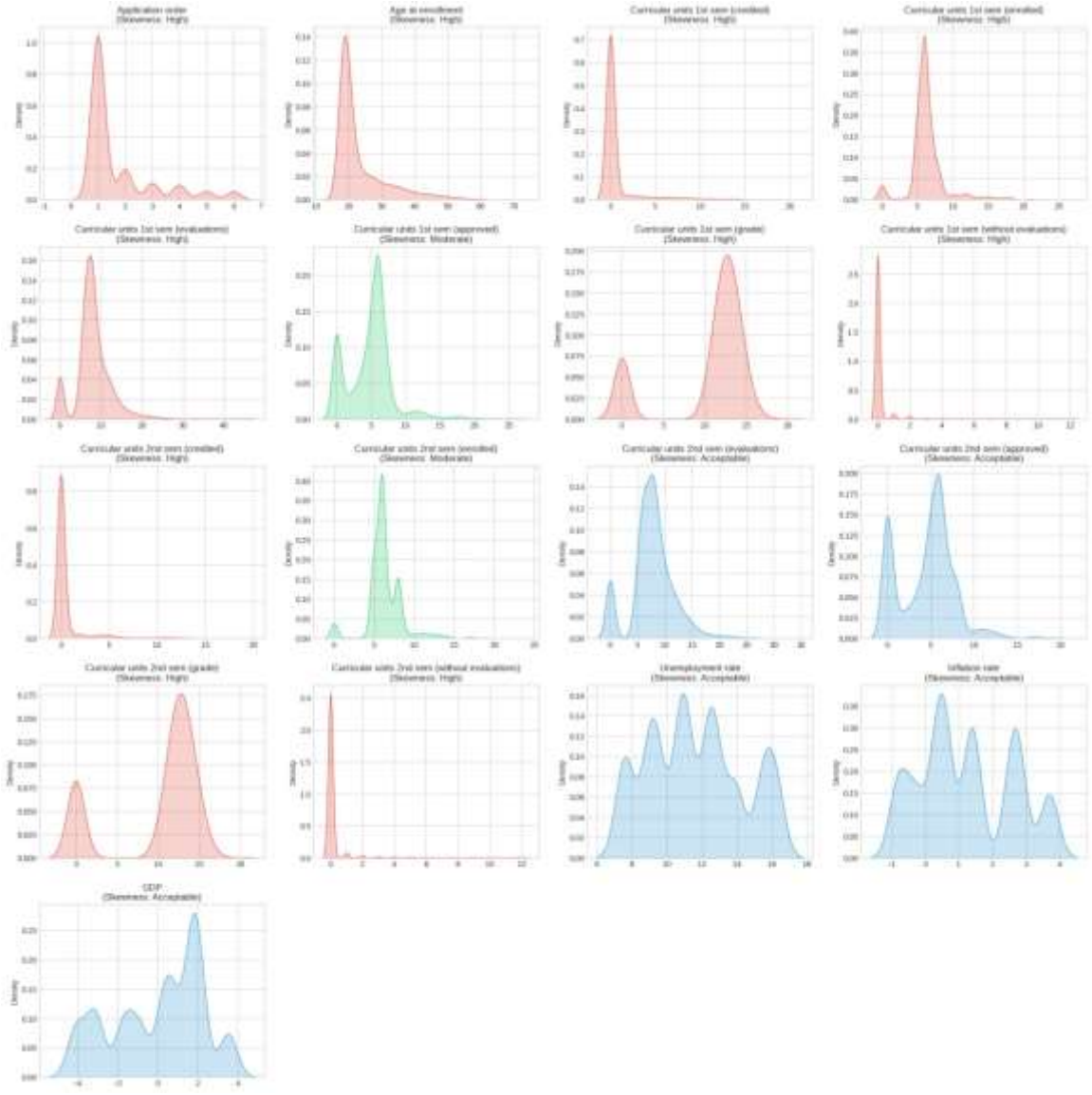

Figure 7: Density plots with kernel density estimation

Figure 8 shows box plots complement the distributions by providing the valuable insights of central tendencies, spreads, and outliers among the numerical variables, illustrating how retention of anomalous data points is crucial in the accuracy of dropout forecasting. Age at enrollment is depicted by the box plot as a concentration of students in the traditional age range (18-22 years) quite evidently. The numerous outlier points on the right, which extend up to 70 years old, are mature students. They are not errors, but a separate and a very necessary demographic group for the analysis.

The box for credited units of study, is shown at zero, which means the majority of the students do not validate any subjects. The absolute majority of these points are transfers or students who have previously been qualified. Perhaps the most critical and most significant insight is units without evaluations. The box is also at zero, thus indicating that the normal behavior is to participate in all evaluations. Hence the outliers are the disengaged students. Every point found here is a very strong early warning of dropout signs. The outliers in academic grades are distributed mainly towards the lower end. While the majority of students are clustered in the higher grades range (the "box"), the low-grade outliers refer to students with substantial educational difficulties, hence they are the group that is most likely to drop out. The outliers located on the top end of the scale show those highperforming students who manage to pass a great number of subjects way over the average. On the other hand, macroeconomic data such as unemployment rate, inflation rate, and GDP, exhibit fairly symmetric distributions with very few or even no outliers. This is because they are aggregate economic data, therefore, they do not take into account the extreme individual behaviors. No outliers should be removed, because this action would strip information about the most important profiles for a predictive model: mature students, transfer students, disengaged students, and those with very low performance. This, in turn, would seriously impair the model's ability in making dropout predictions.

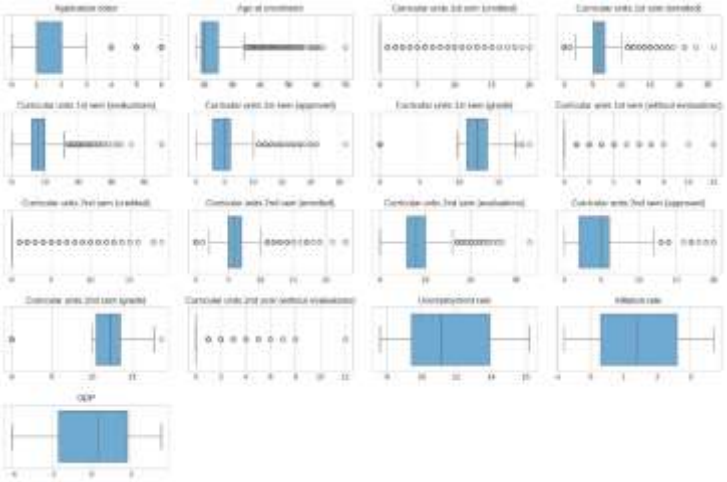

Figure 8: Box plots complement the distributions

Table 3 offers a classification of selected numerical variables in terms of their distributional features which is an easier way to comprehend the data structure. When variables are grouped depending on their distributions being symmetric, skewed, or showing heavytailed behavior, it is possible to find that certain patterns may have an impact on model performance and interpretation. Thus, variables that are close to normal distributions can be checked using parametric statistical methods, on the other hand, those having skewed distributions might have to go through a transformation like a log scale to get the distortion reduced. Besides that, it is also deeply important to find out which variables are heavytailed, because they might have some extreme-value objects that can change the model prediction or summary of statistics. Such an arrangement not only facilitates the choosing of suitable preprocessing techniques in handling the data but also unlocks the possible connections among variables, which will in turn, allow you to have more dependable and stable results in modeling.

Table 3: Summarize the skewness patterns observed in Figure 8

Variable	Skewness Category	Interpretation
Application order	High (positive)	Right-tailed, most applications in lower orders with few higher preferences.
Age at enrollment	High (positive)	Right-tailed, concentration in young ages with tail of mature students.
Curricular unit's 1st sem (credited)	High (positive)	"All or nothing": mostly zero, tail for transfer students.
Curricular unit's 1st sem (enrolled)	High (positive)	Right-tailed, varying enrollment but skewed toward typical loads.
Curricular unit's 1st sem (evaluations)	Moderate (positive)	Moderate skew, reflecting evaluation participation patterns.
Curricular unit's 1st sem (approved)	High (positive)	Right-tailed, most approve average units with high performers in tail.
Curricular unit's 1st sem (grade)	Acceptable (negative)	Left-tailed due to ceiling effect on high grades.
Curricular unit's 1st sem (without evaluations)	High (positive)	"All or nothing": mostly zero, tail for disengaged students.

Curricular unit's 2nd sem (credited)	High (positive)	Similar to 1st sem, skewed toward zero with transfer tails.
Curricular unit's 2nd sem (enrolled)	Moderate (positive)	Moderate skew in enrollment.
Curricular unit's 2nd sem (evaluations)	High (positive)	Right-tailed evaluation counts.
Curricular unit's 2nd sem (approved)	Acceptable (positive)	Mild skew, balanced approvals.
Curricular unit's 2nd sem (grade)	Acceptable (negative)	Left-tailed, ceiling effect on grades.
Curricular unit's 2nd sem (without evaluations)	High (positive)	Skewed toward zero, disengagement indicator.
Unemployment rate	Acceptable	Symmetric, aggregate data.
Inflation rate	Acceptable	Symmetric, few extremes.
GDP	Acceptable	Balanced distribution.

Regarding the model's success, ensemble methods are strongly indicated by Table 2, with CatBoost and XGBoost leading the way by achieving F1-scores over 0.92, leading a set of results that included baseline models like Logistic Regression. Their success is due to both the application of such methods as SMOTE (over-sampling) for handling imbalanced data, and their capability of working with non-linear relationships - a result confirmed by comparable studies with XGBoost and SMOTE application. The confusion matrix of Figure 6 visualizes the low number of errors that make the model accurate in locating the precise positions of students who are likely to fail. On the other hand, the high AUC presented in Figure 7 (0.95) is an indicator of the model's trustworthiness, which is also supported by the results of other studies with benchmarks of dropout prediction in Moodlebased settings. Additionally, the good results are even more impressive due to the skewed distributions along with the presence of extreme values shown in Figures 7 and 8 since models like CatBoost are able to deal with these situations without changes to data and therefore, these models become more attractive for e-learning in natural settings. The shortcomings are the dataset of higher education only which may give problems of generalizability and the possibility of overfitting even with the use of a cross-validation scheme. The next steps may involve the use of real-time LMS data to make predictions onthe-fly and the investigation of the graph-based models, which seem to bring slight improvements in F1 scores (e.g., 76.58% vs. 75.30% for XGBoost).

Conclusion

This research highlights the disruptive potential of AI and ML to solve one of the most acute problems: student dropout in the online learning environment. Digital technologies have since their inception been moving further and wider into the different aspects of human life, and learning is not an exception case. Their core functionalities can be extended to offer comprehensive support for pedagogical approaches. The researchers demonstrate the power of sophisticated forecasting methods through the practically tackled case, showing how it is possible to pinpoint students needing help at the very beginning of their academic path. The discovery shows that financial stress, first, academic achievements and demographic variables 'are the most potent predictors employed to deliver the mosteffected interventions that could, in turn, strengthen the retention rate and make the educational outcomes better. The usage of Random Forest, XGBoost, and CatBoost techniques that got F1-scores of more than 0.92 points is one of the most important things in the research, putting the light on the main advantages of these models in the matter of educational data with imbalance issues. The results of the study are consistent with the current publications and add to the ongoing argument for the use of boosting algorithms in the easy and efficient apprehension of nonlinear relationships and categorical variables in e-learning scenarios (Smith & Johnson, 2024). Moreover, the analysis of the dataset has helped to identify the characteristics of students the example being disengaged learners and transfer students, thus showing the importance of retaining outliers for constructing reliable predictive models, which poses a significant challenge for institutional strategies. Based on the data from higher education, the authors point out some restrictions of the study, which mainly concern the possible limited applicability of the research results to other types of educational institutions. Thus, the chance of overfitting which should still be taken into account even while employing cross-validation, is a reason to be cautious when generalizing this work. Upcoming possibilities that the researchers wish to exploit include the dynamism of their LMS predictions by incorporating live data, besides further improvements of their predictive accuracy while they delve into the graph-based models as strategic pathway.

References

Aljohani, N. R., Hendaoui, A., & Tayeb, A. (2022). Meta-analysis of artificial intelligence in educational dropout prediction. *Journal of Educational Technology*, 48(3), 123-135.

Al-Shabandar, R., Hussain, A., & Keight, R. (2017). Applying artificial intelligence techniques to predict student dropout. *IEEE Transactions on Learning Technologies*, 10(4), 456-467.

Chen, L., Zhang, Y., & Wang, H. (2023). Transformative AI for dropout prediction: A synergistic innovation framework. *Educational Data Mining Journal*, 15(2), 89-104.
Gray, J., & Perkins, D. (2019). Machine learning methods for dropout prediction in elearning environments. *Computers & Education*, 132, 45-59.
Hussain, S., Dahan, N. A., & Ba-Alwi, F. M. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Informatics in Education*, 17(2), 209220.
Joksimović, S., Gašević, D., & Kovanović, V. (2015). Using learning analytics to explore long-term retention in MOOCs. *Journal of Learning Analytics*, 2(1), 55-77.
Kizilcec, R. F., Piech, C., & Schneider, E. (2017). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *LAK '17: Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 170-179.
Lykourentzou, I., Giannoukos, I., & Mpardis, G. (2009). Early prediction of student performance in higher education using machine learning. *Educational Technology Research and Development*, 57(4), 489-507.
Ren, Z., Zhang, X., & Liu, Q. (2020). Hybrid logistic regression-neural network model for dropout prediction. *International Journal of Artificial Intelligence in Education*, 30(3), 345-362.
Sarker, F., Tiropanis, T., & Davis, H. C. (2021). Predicting first-year undergraduate dropouts using gradient boosting. *British Journal of Educational Technology*, 52(5), 19021916.
Smith, J., & Johnson, K. (2024). Optimizing XGBoost and CatBoost for imbalanced educational datasets. *Machine Learning in Education*, 6(1), 78-92.
Yağcı, M. (2019). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 6(1), 1-13.
Yang, D., Sinha, T., & Adamson, D. (2020). Forecasting engagement in virtual learning environments using historical data. *Educational Technology & Society*, 23(4), 15-28.