Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 7, Special Issue, 2023

"Current Issues, Trends, and New Ideas in Education"

Educational Data Mining to Predict Bachelors Students' Success

David Jacob [1], Roberto Henriques [2*]

[1] *NOVA School of Business and Economics (NOVA SBE), Carcavelos, Portugal.*

[2] *Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, Lisboa, Portugal.*

Abstract

Predicting academic success is essential in higher education because it is perceived as a critical driver for scientific and technological advancement and countries' economic and social development. This paper aims to retrieve the most relevant attributes for academic success by applying educational data mining (EDM) techniques to a Portuguese business school bachelor's historical data. We propose two predictive models to classify each student regarding academic success at enrolment and the end of the first academic year. We implemented a SEMMA methodology and tried several machine learning algorithms, including decision trees, KNN, neural networks, and SVM. The best classifier for academic success at the entry-level reached is a random forest with an accuracy of 69%. At the end of the first academic year, an MLP artificial neural network's best performance was achieved with an accuracy of 85%. The main findings show that at enrolment or the end of the first year, the grades and, thus, the student's previous education and engagement with the school environment are decisive in achieving academic success.

Keywords:

Academic Success;
Student Success;
Educational Data Mining;
Machine Learning.

Article History:

Received:	27	January	2023
Revised:	04	July	2023
Accepted:	16	July	2023
Published:	27	July	2023

1- Introduction

At every university, a student's performance is critical because it influences academic accomplishment, which is one of the essential factors in assessing the institution's overall excellence [1]. The quality of the educational system, and therefore academic success, is perceived as the most critical factor in countries' economic and social development. In contrast, academic failure has significant negative social effects [2]. Technology has revolutionized data collection, analysis, and interpretation, and it has influenced practices, processes, and decision-making in various fields, and the field of education is no exception [3]. As higher education schools increasingly gather data on candidates, students, and graduates over the years, it becomes more feasible to use data mining to find hidden patterns to predict and understand the fundamental characteristics of the student population's academic and learning success [4]. In this context, a recent discipline emerged, educational data mining (EDM), which devotes itself to developing the techniques to explore this increasing amount of data from higher education institutions (HEI) to better understand students' behaviour [5, 6]. EDM has emerged as a powerful tool for educators to anticipate scenarios such as disengagement from coursework or dropping out of school. It also allows analysing internal factors using statistical methods to predict students' academic performance [7]. The ability to predict student performance and identify students at risk of failure is an expanding research area [8, 9].

Although college students' final exam scores partially reflect their learning effects, absolute scores have limitations in evaluating the learning situation. Variations in course difficulty and marking standards across different teachers may affect the absolute scores' accuracy. To ensure talent quality, colleges and universities should not solely rely on scores

[*] **CONTACT**: roberto@novaims.unl.pt

DOI: http://dx.doi.org/10.28991/ESJ-2023-SIED2-013

for student evaluations. Instead, they should also analyse learning effects, predict academic performance based on the analysis results, and issue academic warnings promptly.

In this paper, we focus on using EDM techniques on a Portuguese business school bachelor's data spanning from the academic year 2007/2008 to 2017/2018 to understand which features are more relevant to academic success at entry and the end of the first enrolment year. These models classify students as successful when they graduate within three enrolment years and unsuccessful otherwise. We used the proposed approach by Hampton [10], which consists of step-by-step knowledge discovery in databases (KDD). After the data set preparation and stratification, we performed feature selection using stepwise logistic regression. Then we implemented several learning algorithms to model the problem, and finally, we assessed the predictions through various performance measures.

This paper's research problem is to build prediction models to evaluate students' academic achievement and predict future achievement objectively. To accomplish this, the study formulates the following two research questions:

- *Can we predict the students' academic success at the enrolment moment?*
- *Can we improve students' academic success prediction after the first year of attending the program?*

We propose modelling academic success in two moments for anticipated support for decision-makers. The first refers to enrolment at the beginning of the first academic year; thus, no internal information exists regarding the student in the program. The second moment refers to the end of the first academic year, where data regarding students' behaviours during the year is collected and included in the analysis.

2- Related Work

Globally, academic success has received more attention in recent decades as governments have realized that students require high achievement to thrive in a global environment [11]. More interest is then shown in discovering the factors contributing to academic success because doing so might help lower the high rates of academic failure [12]. Regarding educational literature, academic success is a term that has been widely used in conjunction with higher education and serves as an umbrella for various student outcomes. However, many perceive the degree's conclusion as its ultimate indicator. More traditional measures of academic success in higher education include various indicators that comprise college grades, graduation, the graduate record examination for postgraduate admissions, and performance on examinations such as the certified public accountant examination tests. Other traditional views even refer to other measurable indicators such as graduates' employment and salary [9, 13].

Mentkowski & Austin [14] presented the Input-Environment-Outcome framework to assess academic success. This conceptual model suggests that the academic success outcome (dependent variable) relies exclusively on the inputs and environment (independent variables). This framework was later reviewed and explained in Terenzini and Reason [15], which clarified that the inputs concerned student characteristics such as socio-demographic, academic, and social background, while the environment referred to factors educators could control, such as curricula and policies of the HEI. Finally, the outcomes translate into the knowledge, skills, and other competencies students concluding the programs acquire. We should note that the inputs relate to the outcomes and the environment, meaning that the environment serves as a mediator, and its relationship with the outcomes reflects the inputs' influence.

Tinto [16] stated that the greater the student's involvement in the academic and social life on campus, the greater their likelihood that they would persist and achieve academic success. Persistence is put in a prominent place and is seen as the driving force that leads to the ultimate achievement of student success. The student's involvement with the higher education system, particularly with their peers and faculty, is deemed intrinsically related to student retention and should be the institutions' core concern [17]. Institutions must develop and employ programs to enhance the persistence of all their students and make these programs endure over time.

For Kuh et al. [9], the pre-college experience, such as family support, academic preparation, and financial aid, was the preparation phase to succeed in higher education. The authors considered student engagement a prominent and central position of the HEI experience and comprised the student behaviours and institutional conditions. Student engagement is usually translated into contact with faculty or their peers' cooperation, as well as active learning. The more that students engage in social and academic activities, the more likely they will persist and graduate. Student engagement directly connects to student behaviour and the time dedicated to academic activities, but also to institutional conditions such as deploying resources or organizing the programs and other activities that generally lead to persistence and subsequent graduation.

EDM consists of the techniques applied to explore the unique characteristics of higher education data [18]. EDM benefits HEIs by uncovering insights from hidden student data patterns. The models originated by applying data mining techniques that provide decision-makers with the necessary information to understand the main features that influence the students' performance. This way, they can anticipate preventive measures that will positively affect the learners' academic path. This will result in a competitive advantage for the HEI because it might translate into higher retention and the students, who may improve their academic performance and learning experience.

Fernandes et al. [19] developed a model for predicting students' academic achievement that incorporated demographic characteristics and in-term activity grades. The model used gradient boosting machine (GBM) classification models and found that the best indicators of achievement scores were the previous year's scores and unattendance, as well as demographic characteristics such as neighbourhood, school, and age. The authors suggested that the model could inform the development of policies to prevent failure. In 2017, Hoffait and Schyns [20] introduced a model that utilized students' prior academic achievements to forecast their success in upcoming semester courses. Rebai et al. [21] also proposed a machine learning-based model for identifying key factors affecting school performance and their relationships. Their regression tree analysis revealed that school size, competition, class size, parental pressure, and gender proportions were the most important factors associated with higher performance, while the random forest algorithm results showed that school size and the percentage of women had the greatest impact on model accuracy.

Musso et al. [22] developed a machine learning model to predict academic performance and dropouts based on learning strategies, social support, motivation, demographics, health, and academic performance characteristics. They found that learning strategies had the highest effect on predicting GPA, while background information had the greatest effect on determining dropouts. Waheed et al. [23] used artificial neural networks to design a model based on students' records of their navigation through the learning management systems (LMS), which showed that demographics and clickstream activities had a significant effect on student performance and the deep learning model could be a useful tool for early prediction of student performance.

Xu et al. [24] analysed the relationship between Internet usage behaviours of university students and academic performance and predicted students' performance using machine learning methods. They found that Internet connection frequency features were positively correlated with academic performance, while Internet traffic volume features were negatively correlated. Bernarcki et al. [25] investigated whether log records in the learning management system alone would be enough to predict academic achievement. They found that the behaviour-based prediction model successfully predicted 75% of those who would need to repeat a course and identified students who might be unsuccessful in subsequent semesters for support. Finally, using mostly demographic variables, the work proposed by Cruz-Jesus et al. [26] applied machine learning techniques such as random forest, logistic regression, k-nearest neighbours, and support vector machines to achieve a predictive accuracy ranging from 50% to 81%.

According to the literature review, predicting students' academic performance and providing support to those at risk are crucial for improving the quality of education. Previous studies have used various variables, including digital traces [19, 23, 24], demographic characteristics [21, 25], learning skills, study habits, and academic performance characteristics [22], to predict performance. Most models have achieved prediction accuracy ranging from 70% to 95%, but collecting and processing such diverse data requires significant time and expertise. Also, relying solely on such data may not always provide an accurate means of preventing academic failure.

The study concerns predicting students' academic achievement using only data available from the national application and data collected from the students' information system during the first academic year.

3- Research Methodology

The EDM field focuses on developing methods for exploring the unique data types of educational environments [27]. We used the SEMMA methodology [10] in this research. SEMMA stands for sample, explore, modify, model, and assess, and the SAS Institute developed it. It can be seen as a practical implementation of the five stages of the KDD process [28]. We chose SAS Enterprise Miner [29] because it encloses a set of tools that can support the whole data mining process, from data sampling to model assessment.

Figure 1 presents the steps taken in this proposal. The sample step consisted of extracting a relevant sample from the business school ERP, extracting a sample of records with the relevant attributes that could answer our problem. In the explore step, we used the available data visualization tools to understand our data set and search for trends and anomalies in the data. We modified and created variables during the modification step to prepare the data set for the modelling phase. The model phase consisted of applying learning algorithms to the available data sets to create models that could classify students in what concerns their academic success. The last step was assessing the different models' performance and deriving the main insights and conclusions from the results.

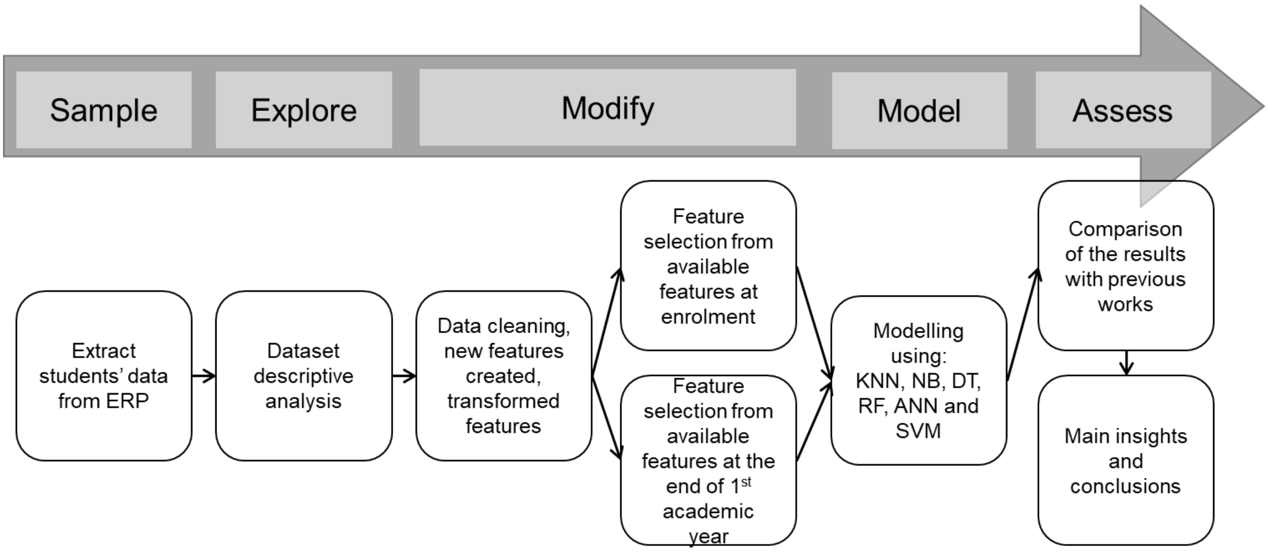
Sample
Explore
Modify
Model
Assess
Extract students' data from ERP
Dataset descriptive analysis
Data cleaning, new features created, transformed features
Feature selection from available features at enrolment
Feature selection from available features at the end of 1st academic year
Modelling using: KNN, NB, DT, RF, ANN and SVM
Comparison of the results with previous works
Main insights and conclusions

Figure 1. **SEMMA methodology steps**

3-1- Sample and Explore

The HEI ERP, which the academic services manage, encloses students' data from enrolment until graduation. It stores the curriculum data of the enrolled students and their personal, professional, and academic information in a relational database that can be queried through the structured query language.

The data extracted spans the academic year 2007/2008 to 2017/2018. For each student, they contained the available socio-economic, socio-demographic, previous education, the European Credit Transfer and Accumulation System (ECTS) completed per academic year, and the grades obtained during the first enrolment year.

A preliminary analysis showed that 53.3% of the students were enrolled in the bachelor's in management, while the remaining 46.7% were enrolled in the bachelor's in economics. Concerning the gender balance, 49.6% of all enrolled students were female, while the remaining 50.4% were male. Most of the students (92.2%) were enrolled in their first choice of program and HEI, and most of their entry types (84%) were "general enrolment". More than 99.2% of the records were of students who, at enrolment, had only a high school degree; 96.1% studied previously in Portugal; and 96.6% were Portuguese. The average high school GPA was 171 (on a scale of 0 to 200), and the average grade of the mandatory math exam was 177. Regarding the parents' education, 60.2% of the students' mothers had a higher education degree, but only 54.1% of their fathers attained this level of education. Only 7.8% worked while studying, and 6% received social aid scholarships. Regarding academic success, 55% of the students graduated within three years of enrolment.

Considering our goal was to create two models for two moments, the enrolment at the beginning of the academic year and the end of the first academic year, we prepared the data accordingly. For this purpose, we created two data sets: DS_Enrolment and DS_End_Year_1. DS_Enrolment refers to the beginning of the first academic year and only includes the features available at the enrolment stage (Table 1). The data set DS_End_Year_1 consists of features available at the end of the first academic year, including those from DS_Enrolment (Table 1).

This division considers the literature review supporting the fact that student outcomes, namely their academic success, are based not only on their pre-college experiences but also on their engagement with the higher education environment, especially during the first year of enrollment [9, 30]. Such involvement with academic life increases the chances of student persistence, which ultimately leads to academic success [16].

3-2- Data Pre-Processing (Modify)

Considering data quality plays a significant role in a data mining study, pre-processing activities are fundamental to reduce the noise in the data set and create new attributes relevant to our research. The first step in data preparation consisted of data cleaning to fill in the missing values, whereas deducting them based on other variables, with the most common value, or with a constant. The second step was to correct the erroneous values. The third step of the data modification phase consisted of filtering outliers and other unwanted values that could create potential bias when applying the learning algorithms. The fourth step was one of the most important. Some nominal variables presented too many classes; thus, their usefulness was compromised. Therefore, we created dummy variables while the original attributes were dropped.

After these data modification activities, the final data set contained 4,546 records spread across 36 input variables presented in Table 1, 12 of which can only be fully available after the first year's conclusion.

Table 1. Features used

Attribute	Type	Range	Available	Description
age_enrollment	Integer	[16, 38]	On enrolment	Age at enrolment
flag_away_from_home	Boolean	T/F		The address is out of the area
flag_enrollment_CEsp	Boolean	T/F		Special enrolment
flag_enrollment_CG	Boolean	T/F		General enrolment
flag_enrollment_REsp	Boolean	T/F		Other special enrolments
flag_father_higher_edu	Boolean	T/F		Father has higher education
flag_father_working	Boolean	T/F		Father is employed
flag_gap_year	Boolean	T/F		Student took a gap year
flag_gender_male	Boolean	T/F		Gender is male
flag_marital_status	Boolean	T/F		Student is single
flag_max_ed_high_school	Boolean	T/F		Prior higher education studies
flag_mother_higher_edu	Boolean	T/F		Mother has higher education
flag_mother_working	Boolean	T/F		Mother is employed
flag_nationality_portuguese	Boolean	T/F		Portuguese nationality
flag_previous_education_PT	Boolean	T/F		Previous education in Portugal
flag_scholarship_merit	Boolean	T/F		Received merit scholarship
flag_scholarship_social_aid	Boolean	T/F		Received social aid scholarship
flag_special_need	Boolean	T/F		Student has special needs
flag_special_support	Boolean	T/F		Student has special support
flag_student_not_working	Boolean	T/F		Working student
preference_order	Integer	[1, 6]		Program and HEI preference order
hs_grade_candidacy	Integer	[0, 200]		Candidacy grade
hs_grade_GPA	Integer	[0, 200]		High school GPA
hs_grade_math_exam	Integer	[0, 200]		Math exam grade
he_ECTS_approved	Integer	[0,60]	After first year's conclusion	ECTS approved in the first year
he_flag_1st_year_60_ECTS	Boolean	T/F		Completing 60 ECTS in the first year
he_grade_calculus_I	Integer	[0, 20]		Calculus I grade
he_grade_calculus_II	Integer	[0, 20]		Calculus II grade
he_grade_data_analysis_prob	Integer	[0, 20]		Data analysis and prob. grade
he_grade_fin_accounting	Integer	[0, 20]		Financial accounting grade
he_grade_law_eco_business	Integer	[0, 20]		Law for econ. and bus. grade
he_grade_linear_algebra	Integer	[0, 20]		Linear algebra grade
he_grade_princ_macroeco	Integer	[0, 20]		Macroeconomics grade
he_grade_princ_management	Integer	[0, 20]		Management grade
he_grade_princ_microeco	Integer	[0, 20]		Microeconomics grade
he_grade_statistics_eco_mng	Integer	[0, 20]		Statistics grade

3-3- Predicting Students' Success at their Entrance

The variables' importance assesses the usefulness of each attribute in predicting the student success classification. We calculated the variable worth using the Gini split worth statistic generated by building a decision tree of depth 1 [29]. Concerning the data set DS_Enrolment, Figure 2 shows that the best predictor of academic success at the beginning of the first enrolment year is the high school GPA. It is followed by the math exam grade, enrolment age, general enrolment type, and if the previous education was from Portugal.

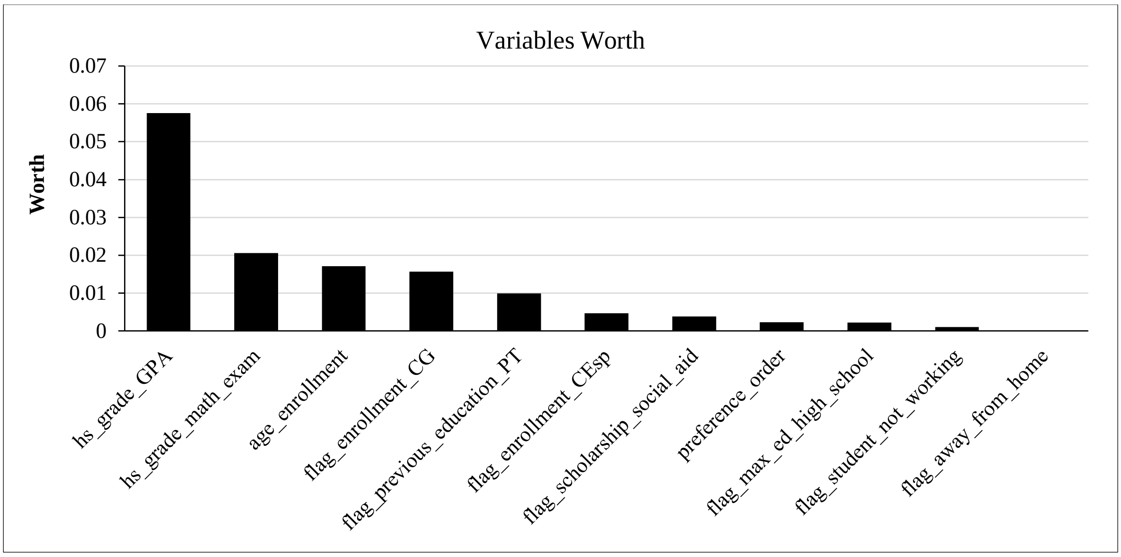

Figure 2. DS_Enrolment variables worth

Looking at the essential variable, GPA (Figure 3), we can see that, for students with high school grades between 120 and 140, only 14% finished their bachelor's degree within three years. Concerning students with a high school GPA between 140 and 160, only 32% achieved academic success, a value that grew to 52% when the GPA fell between 160 and 180. Of students with a GPA between 180 and 200, 76% succeeded. Concerning the math exam grade, we can conclude that the more successful students tended to have a math grade above 140. Considering their age at enrolment, we can observe that the older the student at entry, the less they managed to finish their bachelor's degree within three years. Finally, 58% of the students whose enrolment type was the general one finished their degree within three years, while for other enrolment types, only 30% completed the program within three years.

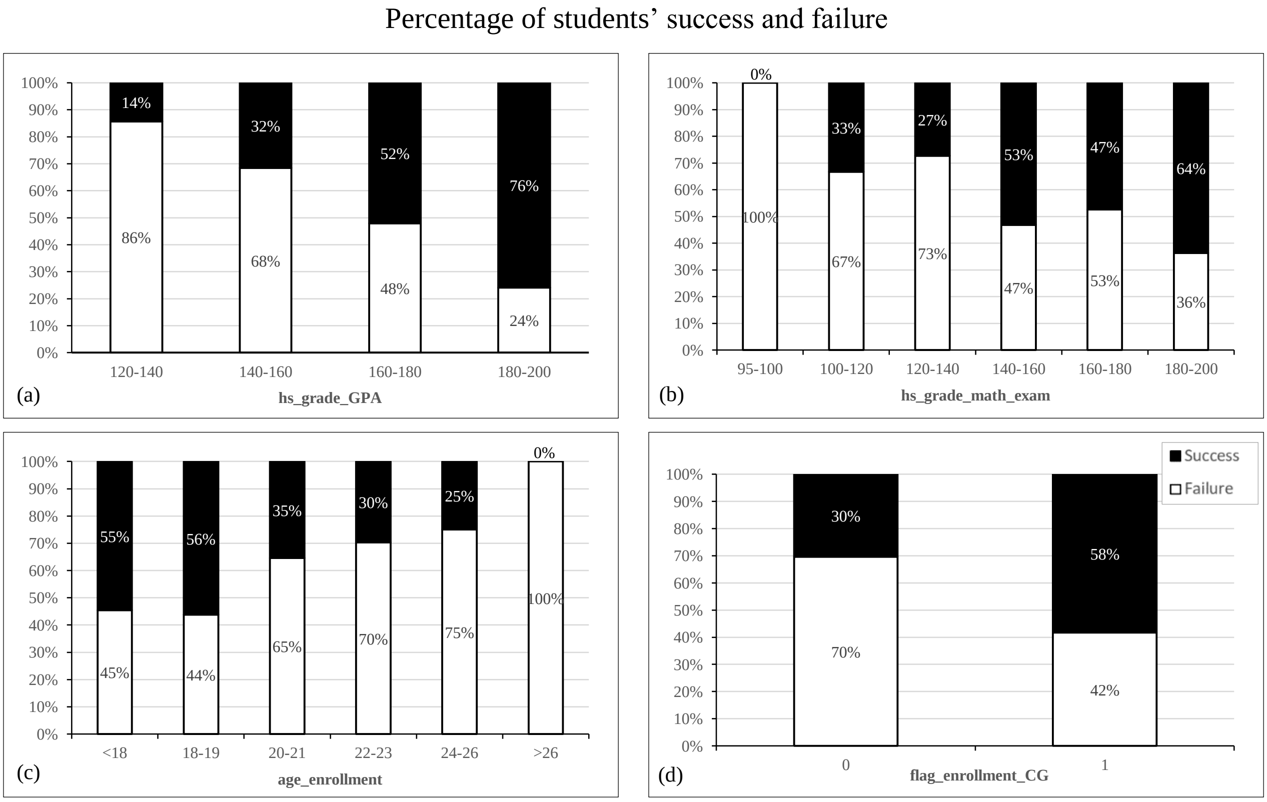

Figure 3. Percentage of students' success and failure by: a) GPA, b) math exam grade, c) age at enrolment, and d) type of enrolment

3-4- Predicting Students' Success at the End of the First Year

Figure 4 presents the variables' worth regarding the data set DS_End_Year_1. Without surprise, we can see that features collected during the first academic year are better predictors when compared to those gathered at the beginning of the academic year.

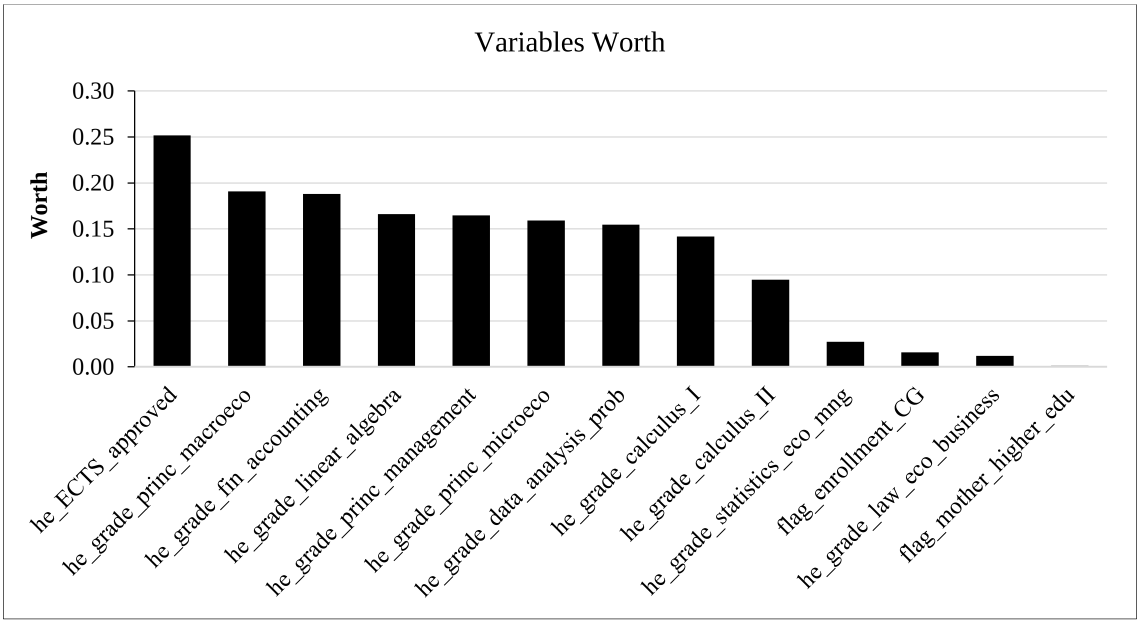

Figure 4. **DS_End_Year_1 variable worth**

At the end of the first enrolment year, the most important variables to predict the target are the number of ECTS completed, followed by the grade in Principles of Macroeconomics, Financial Accounting, and Linear Algebra.

From the number of ECTS approved at the end of the first enrolment year, we can observe that the students who completed less than 30 ECTS did not manage to achieve academic success (Figure 5). For students completing between 30 and 40 ECTS, 8% achieved academic success, and for students completing between 40 and 50 ECTS, 43% achieved success. Of the students who completed over 50 ECTS during their first year, 81% finished their bachelor's degree within three years.

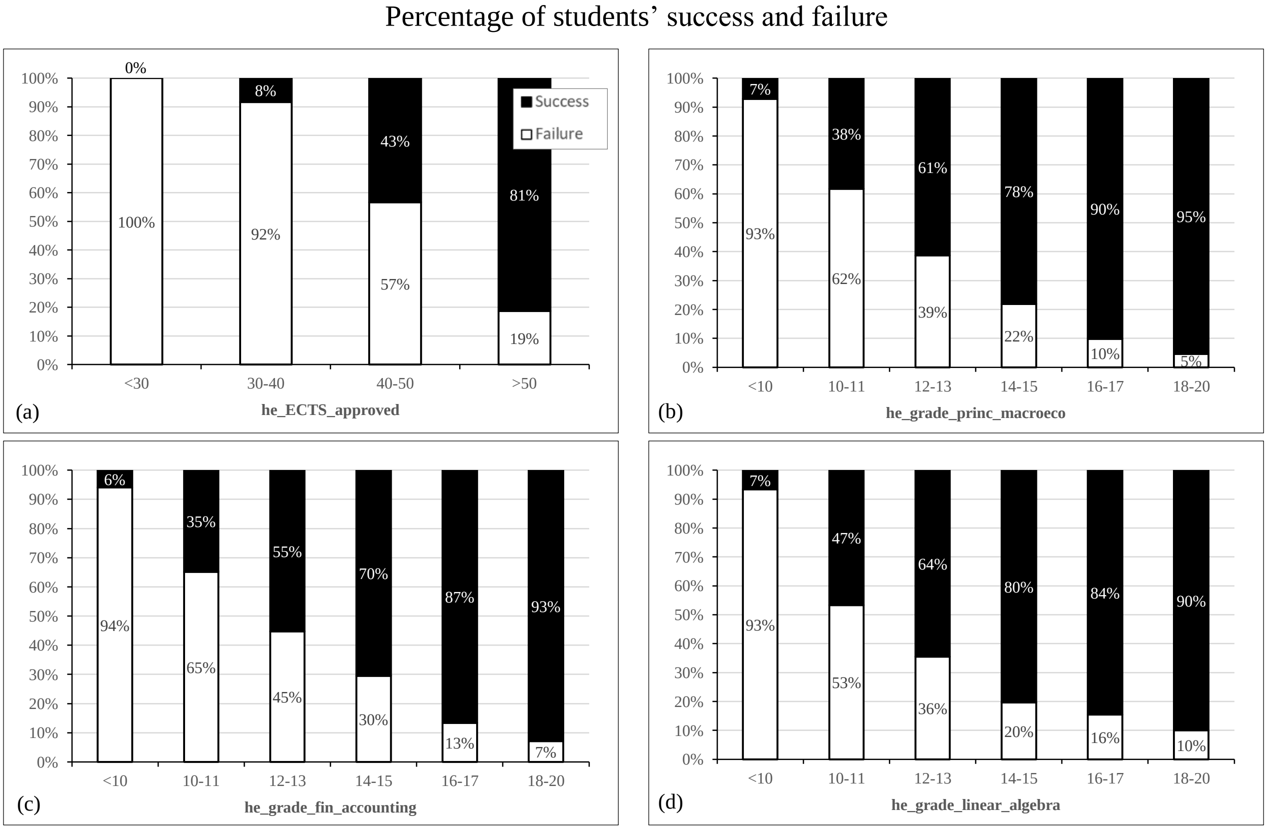

Figure 5. **Percentage of students' success and failure by a) the number of ECTS completed, b) the grade in Principles of Macroeconomics, c) the Financial Accounting grade, and d) the Linear Algebra grade**

Regarding the grade of Principles of Macroeconomics, we can see that most students below 12 failed to achieve academic success. The same applies to Financial Accounting grades and Linear Algebra grades.

3-5- Modelling

Each data set was partitioned into three stratified subsets. The training set (70% of the data), the validation set (15% of the data), and the test set (with the remaining 15%).

We determined the most relevant features through logistic regression for stepwise selection. The purpose of variable selection is to feed the learning algorithms with the most relevant features and avoid certain constraints, such as the curse of dimensionality, model overfitting, or the use of highly correlated variables. Using a stepwise selection, the chi-square and p-values are computed and inserted into the model according to their highest significance. However, they can be removed afterwards if inserting another variable increases the model's performance [31]. By analysing the maximum likelihood estimates, we retrieved 11 features as the most significant at entry (Table 2) and 13 at the end of the first academic year (Table 3).

Table 2. DS_Enrollment data set. Features are selected using a stepwise selection at the enrolment phase

Attribute	Estimate	Chi-Square	Pr > ChiSq
intercept	−9.7518	36.34	< .0001
hs_grade_GPA	0.0671	258.73	< .0001
hs_grade_math_exam	0.0164	30.53	< .0001
age_enrollment	−0.3211	25.68	< .0001
flag_away_from_home	0.1700	16.02	< .0001
flag_scholarship_social_aid	−0.4686	11.00	0.0009
flag_max_ed_high_school	−1.2694	5.53	0.0187
flag_enrollment_CEsp	−0.6393	17.51	< .0001
flag_enrollment_CG	−0.9273	44.73	< .0001
flag_student_not_working	−0.1726	5.27	0.0217
flag_previous_education_PT	−0.5065	10.52	0.0012
preference_order	−0.2205	12.34	0.0004

Table 3. DS_End_Year_1 data set. Features are selected using a stepwise selection at the end of the first academic year

Attribute	Estimate	Chi-Square	Pr > ChiSq
intercept	−10.9643	464.81	< .0001
he_grade_linear_algebra	0.0865	22.71	< .0001
he_grade_data_analysis_prob	0.0642	10.46	0.0012
he_grade_calculus_I	0.0460	3.89	0.0485
he_grade_calculus_II	0.1014	51.58	< .0001
he_grade_fin_accounting	0.0902	22.14	< .0001
he_grade_law_eco_business	0.0452	12.06	0.0005
he_grade_statistics_eco_mng	0.0530	19.25	< .0001
he_grade_princ_management	0.0990	18.58	< .0001
he_grade_princ_macroeco	0.1126	48.52	< .0001
he_grade_princ_microeco	0.0669	9.85	0.0017
flag_enrollment_CG	−0.2842	9.37	0.0022
flag_mother_higher_edu	0.1482	6.14	0.0132
he_ECTS_approved	0.0695	61.80	< .0001

The learning algorithms used to create the models were k-nearest neighbour (KNN), decision tree (DT), random forest (RF), Bayesian classifier (NB), artificial neural network (ANN), and support vector machine (SVM). Table 4 presents the values of the hyperparameters used for each algorithm.

3-6- Assess

We proposed using metrics based on the confusion matrices to assess the different models' performance, such as accuracy, precision, recall, specificity, F1 score, and AUC. Analysing the performance metrics revealed how well the classifiers predicted academic success and whether the model should be employed for that purpose (Table 5).

Table 4. Learning algorithms parameters

Learning Algorithm	Parameters
KNN	Number of neighbours = 32
DT	C4.5 algorithm
	Maximum depth = 10
RF	Maximum number of trees = 110
	Maximum depth = 10
NB	Network model = Naïve Bayes
ANN	Multilayer perceptron with four hidden layers
	50 iterations
SVM	Linear (no kernel)
	Tolerance = 0.1
	Penalty (C – outside the margin) = 1

Table 5. Performance metrics

Measure	Formula	Description
Accuracy	$(TP + TN)/(Total\ Records)$	Representation of the model's general effectiveness that presents the percentage of records it correctly classifies.
Precision	$TP/(TP + FP)$	Agreement between the positive labels in the data set, when compared with all the records, predicted positively by the model.
Recall (Sensitivity)	$TP/(TP + FN)$	Performance of the model in predicting correctly positive records.
Specificity	$TN/(TN + FP)$	Performance of the model in predicting correctly negative records.
F1 Score	$2 \times (Precision \times Recall)/((Precision + Recall))$	The harmonic mean of precision and recall.
ROC AUC	Plot sensitivity on the y-axis against 1 – specificity on the x-axis at its different thresholds.	AUC represents the degree of separability and tells how much the model can distinguish between classes.

4- Results and Discussion

This study proposes two models based on machine learning algorithms to predict undergraduate students' academic success, taking available data at the first enrolment and end of the first academic moments. Concerning the models' assessment, we show and compare the performance metrics extracted from the classification of the records in the testing set (n = 683) of data sets, DS_Enrolment and DS_End_Year_1. Notably, 375 records are classified as having achieved academic success, while the remaining 308 records are classified as having failed. The first models to assess and compare are the ones obtained from the 11 selected variables in the data set DS_Enrolment. Table 6 shows the confusion matrix of the six models created and the performance metrics extracted from them.

Table 6. Models' performance using data at the enrolment

	TP	TN	FP	FN	Accuracy	Specificity	Precision	Recall	F1 Score	AUC ROC
ANN	296	164	144	79	0.6735	0.5325	0.6727	0.7893	0.7264	0.7290
DT	280	175	133	95	0.6662	0.5682	0.6780	0.7467	0.7107	0.7040
KNN	292	130	178	83	0.6179	0.4221	0.6213	0.7787	0.6911	0.6630
NB	283	179	129	92	0.6764	0.5812	0.6869	0.7547	0.7192	0.7160
RF	286	186	122	89	**0.6911**	0.6039	0.7010	0.7627	0.7305	0.7420
SVM	301	158	150	74	0.6720	0.5130	0.6674	0.8027	0.7288	0.7370

Looking at these performance metrics, we can conclude that the best model to correctly classify students' academic outcomes at enrolment is the RF, while the worst is the KNN. Although the SVM and ANN both have better performance predicting positive records than the RF does (SVM recall = 0.8027, ANN recall = 0.7893, RF recall = 0.7627), they perform slightly under it in the other metrics, namely, specificity. These models have similar accuracy, precision, F1 scores, and even identical AUC ROC. What differentiates them is that the RF is much better at predicting true negatives, while the SVM and ANN are slightly better at predicting true positives.

Table 7 shows the confusion matrix for each model created, and the performance metrics extracted using the data set DS_End_Year_1.

Table 7. End of first enrolment year models performance

	TP	TN	FP	FN	Accuracy	Specificity	Precision	Recall	F1 Score	AUC ROC
ANN	345	238	70	30	**0.8536**	0.7727	0.8313	0.9200	0.8734	0.9240
DT	338	225	83	37	0.8243	0.7305	0.8029	0.9013	0.8492	0.8740
KNN	343	231	77	32	0.8404	0.7500	0.8167	0.9147	0.8629	0.9220
NB	341	238	70	34	0.8477	0.7727	0.8297	0.9093	0.8677	0.9220
RF	343	228	80	32	0.8360	0.7403	0.8109	0.9147	0.8596	0.9230
SVM	345	235	73	30	0.8492	0.7630	0.8254	0.9200	0.8701	0.9260

The model showing the best performance in correctly classifying the academic outcome of students at the end of the first academic year is the ANN, while the worst is the DT. Nevertheless, we can also determine that all models, except for the DT, have very similar performances among all the different metrics. These results conclude that with the available features, it is possible to predict academic success within the first year of enrolment, whether at the beginning or end of the academic year.

In fact, at the enrolment stage, with the variables present in the data set DS_Enrolment, and using the RF learning algorithm, it is possible to create a model that achieves good predictions (Accuracy = 0.6911; Specificity = 0.6039; Precision = 0.701; Recall = 0.7627; F1 score = 0.7305; AUC ROC = 0.742). This is considered a good predictor because, although it classifies the successful students better than the unsuccessful, it still shows an accuracy of 69.11% at a stage in which a key factor for academic success is still missing, students' engagement with the academic environment.

Furthermore, at the end of the first academic year, the ANN generates a model in which prediction performance increases significantly from the enrolment stage (Accuracy = 0.8536; Specificity = 0.7727; Precision = 0.8313; Recall = 0.92; F1 score = 0.8734; AUC ROC = 0.924). With an accuracy of over 85% and a precision of over 83%, it is considered a good classifier, especially for students who achieve academic success, because it shows excellent results concerning the recall and the area under the ROC curve.

The best models predicted students' academic success with an AUC ROC of 0.742 at first enrolment and 0.9240 after the first academic year. According to this result, this model can predict academic achievement in the future. If students' success is predicted, they can evaluate their working methods and enhance their performance. Additionally, HEIs can implement early procedures to support these students to avoid failure and possible dropout. Although the models predicted very high student success at the end of the first year, we highlight a good prediction capability at the first enrolment moment, enabling students to obtain earlier support.

The study's findings were compared to previous research that used demographic and socio-economic variables to predict students' academic success. From the model proposed by Hoffait & Schyns [20], which utilized students' academic achievements in prior years to predict their performance in upcoming courses, the authors identified that 12.2% of students had a high risk of failing with 90% confidence. In addition, regarding the work of Waheed et al. [23], the authors proposed a model that accurately predicted students' success or failure with 85% accuracy. Finally, the work proposed by Cruz-Jesus et al. [26] predicted students' academic achievement based on income, age, employment, cultural level indicators, place of residence, and socio-economic data with an AUC ROC of 0.75.

5- Conclusion and Future Directions

Our findings demonstrate that it is possible for HEIs, with the available data enclosed in their ERP, not only to select the most relevant variables for academic success but also to create data mining models to predict it. From our research questions, we wanted to check if we could predict the students' academic success at the enrolment moment or if we could improve the prediction in the case of having data regarding the first year. Regarding the first question, at the entry-stage moment, logistic regression assisted in retrieving the essential variables to answer the problem and allowed the creation of the data set DS_Enrolment. Concerning the most relevant variable at the entry stage, the final high school grade, we should pay special attention to students with scores lower than 160 because historical data show that the vast majority underperform regarding academic success. Although the math exam is not as important, it should be considered that, from the available data, we can observe that students with scores under 140 are more likely not to finish their bachelor's degree within three years. Regarding the student's age at the entry stage, we observed that for ranges between 20 and 26 years old, only around 30% achieved academic success, while none over 26 had success. Regarding entry type, students not entering as general historically showed more difficulty completing 180 ECTS within three years than their counterparts did. Furthermore, around 85% of students who did not complete their previous education in Portugal did not achieve academic success. Thus, we showed that it is possible to predict student success based only on the data available at the entrance moment.

Concerning our second research question, if we can improve the prediction at the end of the first enrolment year, a descriptive analysis was also performed after the most relevant variable selection in the data set called DS_End_Year_1. The main conclusions of the current data reveal that students completing less than 40 ECTS during their first year at the HEI were most likely to fail to complete the program within three years. Furthermore, regarding the grades of the different first-year courses, except Calculus II, students who consistently had lower final scores than 12 were unsuccessful. Regarding Calculus II, the students who passed the course generally achieved academic success.

Assessing the predictive models revealed that the best at predicting academic success at entry was the RF (accuracy = 0.6911, specificity = 0.6039, precision = 0.7010, recall = 0.7627, F1 score = 0.7305, AUC ROC = 0.7420), even though the ANN and SVM were remarkably close regarding performance. In contrast, at the end of the first academic year, the best model was the ANN (accuracy = 0.8536, specificity = 0.7727, precision = 0.8313, recall = 0.92, F1 score = 0.8734, AUC ROC = 0.9240), although the KNN, NB, RF, and SVM presented similar performances at predicting student success. Thus, we showed that after the student attends the first academic year, we can build a more accurate model for predicting academic success and identifying which features of students are good proxies for program success.

Regardless of the data set, all models showed less performance at predicting unsuccessful students. Further analysis comparing the models created with the selected attributes of datasets DS_Enrolment and DS_End_Year_1 against all the features available at the two distinct stages of enrolment revealed that the best performance was achieved using fewer features meaning that feature selection is essential to improving the models' performances. Regarding this work's main limitations, because EDM relies heavily on data, we recommended for future developments that more information on students be thoroughly collected, namely, class attendance, class participation, or extracurricular activities, which would reveal, apart from course degrees, other student engagement variables and would increase the model's predictive power. Furthermore, surveying the students would provide the capability to create models to explore different dimensions of academic success, such as satisfaction with the courses and degree or self-fulfilment.

6- Declarations

6-1- Author Contributions

Conceptualization, D.J. and R.H.; methodology, D.J. and R.H.; formal analysis, R.H.; investigation, D.J.; data curation, R.H.; writing—original draft preparation, D.J.; writing—review and editing, R.H.; supervision, R.H. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422. doi:10.1016/j.procs.2015.12.157.

[2] Najimi, A., Sharifirad, G., Amini, M., & Meftagh, S. (2013). Academic failure and students viewpoint: The influence of individual, internal and external organizational factors. Journal of Education and Health Promotion, 2(1), 22. doi:10.4103/2277-9531.112698.

[3] Baek, C., & Doleck, T. (2020). A Bibliometric Analysis of the Papers Published in the Journal of Artificial Intelligence in Education from 2015-2019. International Journal of Learning Analytics and Artificial Intelligence for Education (IJAI), 2(1), 67. doi:10.3991/ijai.v2i1.14481.

[4] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining Big Data in Education: Affordances and Challenges. Review of Research in Education, 44(1), 130–160. doi:10.3102/0091732X20903304.

[5] Mohamad, S. K., & Tasir, Z. (2013). Educational Data Mining: A Review. Procedia - Social and Behavioral Sciences, 97, 320–324. doi:10.1016/j.sbspro.2013.10.240.

[6] Yu, R., Jiang, D., & Warschauer, M. (2018). Representing and predicting student navigational pathways in online college courses. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale, 44, 1-4. doi:10.1145/3231644.3231702.

[7] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 1-19. doi:10.1186/s40561-022-00192-z.

[8] Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. Knowledge-Based Systems, 161, 134–146. doi:10.1016/j.knosys.2018.07.042.

[9] Kuh, G. D., Kinzie, J. L., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). What matters to student success: A review of the literature. National Postsecondary Education Cooperative, Washington, United States.

[10] Hampton, J. (2011). SEMMA and CRISP-DM: Data mining methodologies. Connecticut, United States. Available online: https://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/ (accesed on May 2023).

[11] Haines, R. T., & Mueller, C. E. (2013). Academic achievement: An adolescent perspective. International guide to student achievement. Routledge, Milton Park, United States. doi:10.4324/9780203850398-4.

[12] Morales-Vives, F., Camps, E., & Dueñas, J. M. (2020). Predicting academic achievement in adolescents: The role of maturity, intelligence and personality. Psicothema, 32(1), 84–91. doi:10.7334/psicothema2019.262.

[13] Simms, S., & Paschke-Wood, J. (2022). Academic Librarians and Student Success: Examining Changing Librarian Roles and Attitudes. Journal of Library Administration, 62(8), 1017-1044. doi:10.1080/01930826.2022.2127585.

[14] Mentkowski, M., & Astin, A. W. (1992). Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education. The Journal of Higher Education, 63(6), 717. doi:10.2307/1982058.

[15] Terenzini, P. T., & Reason, R. D. (2005). Parsing the first year of college: A conceptual framework for studying college impacts. Annual meeting of the Association for the Study of Higher Education, November, Philadelphia, United Stastes.

[16] Tinto, V. (1997). Classrooms as communities: Exploring the educational character of student persistence. Journal of Higher Education, 68(6), 599-623. doi:10.2307/2959965.

[17] Tinto, V. (2006). Research and practice of student retention: What next? Journal of College Student Retention: Research, Theory and Practice, 8(1), 1–19. doi:10.2190/4YNU-4TMB-22DJ-AN4W.

[18] Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 40(6), 601–618. doi:10.1109/TSMCC.2010.2053532.

[19] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94, 335–343. doi:10.1016/j.jbusres.2018.02.012.

[20] Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. Decision Support Systems, 101, 1–11. doi:10.1016/j.dss.2017.05.003.

[21] Rebai, S., Ben Yahia, F., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70. doi:10.1016/j.seps.2019.06.009.

[22] Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. Higher Education, 80(5), 875–894. doi:10.1007/s10734-020-00520-7.

[23] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104. doi:10.1016/j.chb.2019.106189.

[24] Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98, 166–173. doi:10.1016/j.chb.2019.04.015.

[25] Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers and Education, 158. doi:10.1016/j.compedu.2020.103999.

[26] Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. Heliyon, 6(6). doi:10.1016/j.heliyon.2020.e04081.

[27] Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies, 23(1), 537–553. doi:10.1007/s10639-017-9616-z.

[28] Azevedo, A., & Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview. IADIS Multi Conference on Computer Science and Information Systems (MCCSIS 200), 22-27 July, 2008, Amsterdam, Netherlands.

[29] SAS Institute Inc. (2011). SAS ® Enterprise Miner 12.1 ® Reference Help (2nd Ed.). SAS Institute Inc, Cary, United States.

[30] Bowman, N. A., & Garvey, J. C. (2022). Theories, findings, and implications from higher education research on student success. In How College Students Succeed, Routledge, 28-50. doi:10.4324/9781003445159-3.

[31] Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine, 3(1), 17. doi:10.1186/1751-0473-3-17.