ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 3, October 2023, pp.27-46

Machine learning prediction and analysis of students' academic performance

Mirza Pasic[1], Ajdin Vatres[1], Faris Ferizbegovic[1], Hadis Bajric[1], Mugdim Pasic[1]
[1] Department of Industrial Engineering and Management, Faculty of Mechanical Engineering, University of Sarajevo, Bosnia and Herzegovina

ABSTRACT

Analyzing students' academic performance is important for evaluating enrollment criteria which establish the standards required for pupils who finished secondary school to gain admission to a higher education institution. The aims of this research were to develop a machine learning prediction Decision Tree classification model and analyze the performance of engineering students based on their performances during secondary school education. The performance of students was analyzed and measured as a binomial response whether students successfully finished the first and the second study years. The developed model examined general success, number of awards obtained at competitions, special awards, average grades in mathematics, physics, and one of the official state languages during secondary school as predictor variables. The number of courses transferred from the first into the second study year and students' GPA obtained during the first study year were added as predictor variables in the analysis and development of a prediction model for the students' performance during the second study year and their enrollment in the third study year. Developed machine learning prediction model showed that for the performance of enrolled students in the first study year general success of students during secondary school is the most important predictor variable, followed by mathematics and physics grades. However, for the performance of the students enrolled in the second study year the most important predictor variable was number of the courses transferred from the first into the second study year, followed by students' GPA obtained during the first study year and general success. Machine learning Decision Tree classification modeling was shown to be an adequate tool for the prediction of the performance of engineering students during the first and second study years.

Keywords: Machine learning, Decision Tree, Enrollment criteria, Engineering students, Study success

Corresponding Author:

Mirza Pasic
Department of Industrial Engineering and Management
Faculty of Mechanical Engineering
University of Sarajevo
Vilsonovo setaliste 9
E-mail: mirza.pasic@mef.unsa.ba

1. Introduction

The frequency at which various statistical and machine learning methods have been used to predict student performance was analyzed in a systematic literature review of 357 articles on predicting students' academic performance (SAP) using various statistical and machine learning methods. It was concluded that 31.3% of authors are using statistical methods such as linear regression and ANOVA with Decision Trees [1]. Clustering methods [1] are less used than classification techniques and clustering is mostly used as a preparation for applying the model. Following this in [2] a further literature survey is performed, shedding more light on what types of models are most frequently used for SAP prediction. The most common models in use are based on Decision Tree, Naïve Bayes, and Rule-Based algorithms, with GPA, gender, age, and marital status as factors.

In [3] methods for predicting student performance are divided into four categories Regression, Clustering, Decision Tree, and Dimensionality reduction. These methods are used to predict many different student performance indicators including course grade or score, grade point average (GPA) or range of GPA, additionally cumulative grade point average (CGPA), semester grade point average (SGPA), course retention or dropout, program or module graduation, and more. For example, in [4]-[6] course grades or scores are predicted, in [7], [8] GPA and additional parameters, in [9]-[11] student dropout, and in [12] program graduation rate, etc.

When analyzing which factors and factor categories are used for prediction, in [1] between 3 and 10 categories are used. Different authors categorized influencing factors like GPA, gender, family background, motivation, educational background, etc., in different categories. Depending on the research authors used different categories in combination with different data mining techniques. In [13] influencing factors are categorized into cumulative grade point average, engagement time, external assessment, extra-curricular activities, family support, high-school background, internal assessment, social interaction network, study behavior, student demographic, and student interest. On the other hand, in [12] influencing factors are divided into activity and course features, demographic features, learning behavior features, student history record and performance, student record and performance in the current course, and other factors.

After defining which statistical and machine learning methods are used, which student performance parameters are being predicted, and which factors for student academic performance (SAP) are used, several research are of note. In [14] nine different parameters for SAP prediction, gender, race and hometown, GPA, family income, university entry mode, grades Malay Language, English, and Mathematics are used. The prediction methods were Decision Trees, Rule-Based models, and Artificial Neural Networks. The obtained accuracy was 67%, 71,3%, and 68.8% respectively. In [15] is investigated which is the best way to predict the final grade of the postgraduate students of Inonina University Informatics Greece taking into consideration gender, age, material status, number of children, occupation, job associated with computers, bachelor, another master, computer literacy and bachelor in informatics. The research tried to find the best algorithm for predicting the final grade. Some of the techniques performed are Decision Trees, Naive Bayes, Rule-Based, and K-nearest neighbors' algorithms. Some of the techniques such as Naive Bayes and K-nearest neighbors showed 100% accuracy (if not overfitted), while others exhibited lesser results, such as Decision Trees at 68.5% accuracy. The developed Rule-Based model achieved a reported 90.9% accuracy. In [16] SAP prediction models are developed using three selected classification methods: Decision Tree, Naïve Bayes, and Rule-Based considering five independent parameters (hometown, family income, university entry mode and even gender and race). The three algorithms achieved an accuracy of 68.8%, 63.3%, and 68.8% respectively.

As students come from diverse backgrounds which can have unplanned effect on their educational and academic achievements a common approach noted in the literature is to group them into similar, sometimes called homogenous groups, using clustering algorithms. This allows for better resource allocation, as students with added educational needs can be identified. The student performance can also be measured within the group allowing for better tracing of their progress. In this manner cluster analysis of students from Malaysia [17], China [18], and Jordan [19] is performed. Various characteristics that can serve for student grouping into homogenous groups are examined in the literature. Students are grouped according to their performance, behavior (presented here through the number of student activities in which they participated), by the results achieved during their studies, their lifestyle habits and study habits. In [17] clustering of students who fall into the B40 (this label is used for people whose income falls into the bottom 40% of the population, who have an average monthly income of less than RM4850) category is performed. Three types of clustering algorithms are used: k-means, BIRCH and DBSCAN. In [18] students from four universities in China are clustered into homogenous groups. Two k-means algorithms are used, the traditional k-means as well as clustering by fast search and find of density peaks (K-CFSFDP) algorithm. In order to compare the results obtained using these two algorithms, three performance measures are used: Silhouette Coefficient (SC), Calinski-Harabasz Index (CHI) and Davies-Bouldin Index (DBI). In [19] a new cluster-based supervised classifier is developed. Clustering techniques are used to divide students into homogeneous groups. Then a separate classification model is built for each cluster. The dataset contained information on students of BAU University in Jordan. Canopy Cluster is used to determine the potential number of homogeneous clusters. After that, the k-means algorithm with three clusters is used. Trait selection is performed using the ant search algorithm. Multi-layer perceptron (MLP), probabilistic Naive Bayes, J48 and meta EMT are used as classifiers.

The created homogenous groups examined in the previous paragraph are not just measures of external similarity. If student performance is considered as the main division between groups, the created homogenous groups can be expanded into a measure of student performance, replacing od adding to existing grading methods. Thus [20] looked at the possibility of including student self-assessment as a form of additional grading input using cluster analysis. Students from Portugal filled out questionnaires with their basic demographic data. In addition, students filled out forms for self-assessment of their grade. The k-means clustering algorithm was used. The sum of the squared errors of the within-cluster is used to estimate the optimal number of clusters. The space from one to ten clusters is searched, resulting in three clusters being chosen. Additional clustering was done using the k-prototypes algorithm with three clusters. Similarly, in [21] clustering algorithms are used to divide students from Columbia into five clusters. Clustering was performed using the Fuzzy C-Means algorithm (FCM). Defuzzification was done using the Takagi-Sugeno-Kang model. This approach made it possible to use the created clusters as a form of more sensitive grading. In [22], this approach is further augmented with the assumption that data generated in a learning environment should be viewed not as stationary data, but as a flow of information. Students are grouped according to whether they passed or failed. A partially supervised clustering algorithm called Dynamic Incremental Semi-Supervised Fuzzy C-Means (DISSFCM) is used to separate the students.

Information gathered about the students can be used not just to determine their groping, but can serve for predictive analytics. The simplest, jet highly informative type of predictive modeling that can be made in this environment is to predict the student drop-out rate. The drop-out rate of student’s form Slovakia [23] and China [24] are analyzed using a number of supervised machine learning algorithms. In [23] the drop-out rate in a Virtual Learning Environment (VLE) is examined. Inputs used in the developed models consisted of the total number of accesses to the course, total points scored on all assessed tasks during the course, points scored during the partial and final test. The correlation between these selected variables and student status was tested using the standard Person correlation. Six different classification models are created: logistic regression (LR), decision tree (DT), Naive Bayes (NB), support vector machine (SVM), random forest (RF), and finally a neural network (NN). Accuracy, precision, response, classification error and F1 measure are used. The final predictions of the different models were compared with the McNemar matching error test. In [24] a binary classifier for the prediction of student drop-out rate during a course is developed. The input variables are the number of blogs the student read during the course, the number of tasks completed, the number of complaints made during the course, the number of responses to complaints, the number of resources the student accessed, the number of posts made during discussions on the course forums, the number of responses to forum posts, and the number of responses posted in the complaints section. A decision tree (DT) algorithm was used to predict student success after completing the course using data after each week of the course.

The predictive models can be further extended from a binary classifier into a model capable of predicting student GPA using various approaches. One such is in [25], were models for predicting GPA of students are developed. In addition, the relationship between individual variables and the students' average is investigated. The data are collected from students of the third year of computer science studies at the Faculty of Management and Informatics of the University of Zilina. Students were divided into two groups, first by gender, and then an equivalent division was made according to the type of previous education. The difference between the mean values of the groups are tested, using the Shapiro-Wilk test, and the parametric t-test or Mann-Whitney test. The correlation between the collected factors and the average was tested using ANOVA analysis. Three types of algorithms were used for the development of regression models: Multinomial Linear Regression, Decision Trees and finally Random Forest. Mean square deviation (MSE) and mean absolute deviation (MAPE) are used as performance measures.

A number of exploratory analysis linking admission criteria [26], standardized test scores [27], and student enrolment strategies [28] with their performance during studies are found in the literature.

In reference [26] the relationship between the admission criteria of students from Saudi Arabia and their academic performance is examined. A linear regression model is created to examine the relationship between three entrance criteria, namely HSGA (high school grade point average), SAAT (English Scholastic Achievement Admission Test) and GAT (English General Aptitude Test), with the student’s average grade after the first year. Precision, response, F1 measure and accuracy are used as performance metrics. In order to predict the results of students during their studies, four different types of binary classification models are used: artificial neural networks (ANN), decision tree (DT), support vector (SVM) and Naive Bayes.

In [27] standardized test scores and high school grades of students from Bahrain, Saudi Arabia, Kuwait, Oman, the United Arab Emirates and Qatar are examined as predictors of their academic performance. The outputs of the created regression model were the GPA of the first-year exams, the GPA of the fourth-year exams, B.Sc. scores (Bachelor of Medical Science exam scores) and the assessment of clinical knowledge in the form of MD exam scores. HSGPA (English high school grade point average), AGU-MCAT (biology, chemistry, physics, and mathematics) test scores and scores from the English language test were used as input variables.

In [28], the impact of various enrolment strategies used by students from Florida on their academic performance is analyzed. Performance is measured through average cumulative GPA, graduation rate, and the so-called DFW rate (a constructed variable that tells how many D, W, and F grades a student has). Hidden Markov model (HMM) with three states, that corresponded to the enrollment strategies the students use is created. Students enroll in the faculty full-time (Full-time enrollment strategy, FES), partially (Part-time enrollment strategy, PES), or depending on their current situation, use what the authors call a mixed enrollment strategy (Mixed enrollment strategy, MES). The authors tested the difference between the distribution of male and female students depending on the enrollment strategy using the Chi-square test. The same test was used when checking whether students show a difference in enrollment strategy depending on ethnicity/race. Finally, the Kruskal-Willis H test was used to verify the existence of a statistically significant difference between enrollment strategy and students' family income. The difference between the average GPA of groups with different enrollment strategies was determined using the Games-Howell test. Categorization of the all examined studies with key information including country where the study was performed, input variables, methods used, type of output generated and reference number is shown in Appendix A, Table 1.

There are different models regarding enrollment criteria at universities. The most commonly used criteria for enrollment at a university are the entrance exam and performance during secondary school. Enrollment criteria play a crucial role in ensuring that universities admit students who possess the essential competencies and potential to thrive in their chosen fields. By admitting individuals who meet certain enrollment thresholds, universities increase the likelihood of these students excelling academically, graduating on time, and being well-prepared for the labor market or further education.

The aims of this research were to develop a machine learning Decision Tree prediction model and to analyze the academic performance of engineering students based on their performances during their secondary school education, and to determine the most important variables available in this research to predict academic performance of the engineering students. This is important because of defining the enrollment criteria, and to ensure that candidates with essential competences are enrolled. Comparing to other similar studies, this study developed different models and used different set of predictor variables to analyze and predict students' performances during the first and the second study year, and relative importances of the predictor variables were calculated. Also, the data from this research were obtained from Bosnia and Herzegovina which has complex constitutional structure and different economic development and educational autonomy of its regions. The educational legal framework is at the level of the entities and the cantons in Bosnia and Herzegovina. This research was conducted at the Faculty of Mechanical Engineering of the University of Sarajevo.

2. Data and methods

In this section, the data sampling and collection procedure is described, and available student information and explored variables are defined. Furthermore, the performance metrics necessary for the model evaluation are presented.

2.1. Sampling and data collection

Data were collected from the Student Service Office of the Faculty of Mechanical Engineering of the University of Sarajevo for the enrolled students during the 2016/17 and 2017/18 academic years. The total sample size was 557 students. Data provided by the Student Service Office were without names of the students. Data contained the following variables:

a) Grades and awards:

- General success (GES) – secondary school general success is obtained by summing up the GPA obtained for each school year,
- Mathematics (MAT) – mathematics average grade during the secondary school,
- Physics (PHY) – physics average grade during the secondary school,
- Language (LAN) – official language average grade during the secondary school,

- Competition awards (CAW) – number of attended competitions and obtained awards,
- Special awards (SAW) - number of special awards obtained.

b) Type of the secondary school:
- Gymnasium (GYM),
- Technical high school (THS),
- College high school (CHS),
- Economics high school (EHS),
- Other high school.

c) Region of the location of the secondary school
- Canton Sarajevo (CS),
- Zenica – Doboj Canton (ZDC),
- Central Bosnia Canton (CBC),
- Bosnian Podrinje Canton (BPC),
- Tuzla Canton (TC),
- Herzegovina – Neretva Canton (HNC),
- Una – Sana Canton (USC),
- Posavina Canton (PC),
- Other.

d) Performance during the first and the second study year at the Faculty
- Courses Transferred (CTR) – number of courses transferred from the first study year into the second study year,
- GPA1 (GPA1) – GPA obtained during the first study year.

e) Response variable
- Performance 1-2 (P12) – whether the student finished the first study year and enrolled in the second study year,
- Performance 2-3 (P23) – whether the student finished the second study year and enrolled in the third study year.

2.2. Data analysis and methodology

Descriptive statistics was used to present fundamental information about the dataset and to indicate where potential correlations between variables can be found. Following that Decision Tree based classification models with a binary response are developed. Decision Tree classification approach is usually suitable for tabular data and majority of similar studies achieved satisfactory results using this approach. After data analysis, it was determined that there was highly non-linear relationship between predictor variables and independent variable for which Decision Tree approach is suitable. Finally, by using Decision Tree approach it is possible to calculate relative importance of predictor variables. Relative importance of predictor variables allows to determine the most important input variables to predict students' academic performance.

In this research accuracy of the prediction model, true positive rate - TPR (sensitivity), false positive rate – FPR (type I error), false negative rate - FNR (type II error) and true negative rate - TNR (specificity) were calculated for both the first and the second study years, along with corresponding confusion matrices for both the training set and the test set. TPR, FPR, FNR, and TNR are defined as follows [29]:
- True positive rate (TPR) — the probability that a student successfully completed the study year is predicted correctly,
- False positive rate (FPR) — the probability that a student failed to successfully complete the study year is predicted incorrectly,
- False negative rate (FNR) — the probability that a student successfully completed the study year is predicted incorrectly,
- True negative rate (TNR) — the probability that a student failed to successfully complete the study year is predicted correctly.

Accuracy is the metric for the evaluation of classification models. Accuracy is actually the ratio between number of the correct predictions and the total number of predictions.

Accuracy, TPR, FPR, FNR and TNR were calculated using equations (1), (2), (3), (4) and (5):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{TN + FP} \tag{3}$$

$$FNR = \frac{FN}{TP + FN} \tag{4}$$

$$TNR = \frac{TN}{TN + FP} \tag{5}$$

where:
TP − true positive,
TN − true negative,
FP − false positive,
FN − false negative.

The relative importance of predictor variables was calculated for the performances during the first and the second study years. The relative importance of the predictor variable is a measure expressed as a percentage that indicates how much improvement a predictor variable offers in comparison to the most important predictor. These relative variable importance values fall within the range of 0% to 100%, with the most important variable consistently holding a relative importance rating of 100%.

3. Results and discussion

Table 2 depicts the number of students enrolled per study year and per type of secondary school, while Table 3 shows percentages of students per study year and per type of secondary high school. From Table 2 and Table 3, it can be seen that almost half of the students enrolled in the first study year (276 or 49.55%) are coming from a gymnasium, followed by 235 (42.19%) technical high school graduates. In the second study year, out of the total number of students enrolled for the first time in the first study year, gymnasium graduates are with dominant number (116 or 62.70%), followed by 56 (30.27%) students who finished technical high school. A similar pattern can be noticed in the third study year, where out of the total number of students enrolled for the first time in the first study year, 81 students (62.31%) with gymnasium background and 39 (30.00%) students from technical high school completed the second study year and enrolled in the third study year. From Table 2 it can be seen that there is a small number of students coming from college high school and economics high school as well as from other high schools.

Table 2. Number of students enrolled per study year and per type of secondary school

	Total	GYM	THS	CHS	EHS	Other
First study year	557	276	235	17	19	10
Second study year	185	116	56	5	7	1
Third study year	130	81	39	2	7	1

Table 3. Percentage of students per study year and per type of secondary high school

	GYM	THS	CHS	EHS	Other
First study year	49.55%	42.19%	3.05%	3.41%	1.80%
Second study year	62.70%	30.27%	2.70%	3.78%	0.54%
Third study year	62.31%	30.00%	1.54%	5.38%	0.77%

Out of the total number of students who were enrolled for the first time in the first study year, 33.21% of students successfully completed the first study year and enrolled in the second study year. Out of the same total number of students who enrolled for the first time in the first study year and enrolled in the second study year only 23.24 % successfully completed the second study year and enrolled in the third study year. However, out of the number of students who enrolled in the second study year 70.27% of students successfully completed the second study year and enrolled in the third study year.

Table 4 shows the number of students enrolled per study year and per region, while Table 5. depicts the percentage of students per study year and per region. From Table 4 and Table 5, it can be seen that more than half of the students (325 or 58.35%) enrolled in the first study year were coming from Canton Sarajevo, followed by 110 (19.75%) students coming from Zenica – Doboj Canton. Out of 185 students enrolled in the second study year, there were 98 (52.97%) students successfully completed the first study year from Canton Sarajevo, while 41 students (22.16%) from Zenica – Doboj Canton successfully completed the first study year and enrolled in the third study year. Out of 130 students enrolled in the third study year, 68 (52.31%) students successfully completed the second study year from Canton Sarajevo, while 30 students (23.08%) from Zenica – Doboj Canton successfully completed the second study year and enrolled in the third study year.

Table 4. Number of students per study year and per region

	Total	CS	ZDC	CBC	PBC	TC	HNC	USC	PC	Other
First study year	557	325	110	37	22	29	16	13	2	3
Second study year	185	98	41	12	13	5	11	4	0	1
Third study year	130	68	30	8	8	4	8	3	0	1

Table 5. Percentage of students per study year and per region

	CS	ZDC	CBC	PBC	TC	HNC	USC	PC	Other
First study year	58.35%	19.75%	6.64%	3.95%	5.21%	2.87%	2.33%	0.36%	0.54%
Second study year	52.97%	22.16%	6.49%	7.03%	2.70%	5.95%	2.16%	0.00%	0.54%
Third study year	52.31%	23.08%	6.15%	6.15%	3.08%	6.15%	2.31%	0.00%	0.77%

Table 6 shows the percentage of students per type of high school who successfully completed the first and the second study years relative to the first-time enrolled students in the first study year. From Table 6 it can be seen that 42.03% of students with a gymnasium background successfully completed the first study year and enrolled in the second study year (P12). Out of the total number of students enrolled in the first study year, 29.35% with a gymnasium background who successfully finished both first and second study year were enrolled in the third study year (S13). Table 6 shows that 23.83% of the students who finished technical high school successfully completed the first study year and enrolled in the second study year (P12). Out of the total number of enrolled students in the first study year, 16.60% of the technical high school students who successfully finished both first and second study year, were enrolled in the third study year (S13).

Table 6. Percentage of students per type of high school who successfully completed the first and the second study year relative to the first-time enrolled students in the first study year

	GYM	THS	CHS	EHS	Other
P12	42.03%	23.83%	29.41%	36.84%	10.00%
S13	29.35%	16.60%	11.76%	36.84%	10.00%

Table 7 shows the percentage of students per type of high school who successfully completed the second study year relative to students who successfully completed the first study year (P23). It can be seen that gymnasium and technical high school students are almost with the same percentage - 69.83% and 69.64% respectively. It can be noted that economics high school students and other high school students have achieved the same results 100% of the time. This is most likely because there was a small number of students enrolled in the second study year from economics high school and other high schools (7 and 1 respectively).

Table 7. Percentage of students per type of high school who successfully completed the second study year relative to students who successfully completed the first study year

	GYM	THS	CHS	EHS	Other
P23	69.83%	69.64%	40.00%	100.00%	100.00%

Table 8 depicts the percentage of the students per region who successfully completed the first and the second study years relative to the first-time enrolled students in the first study year. From Table 8 it can be seen that 30.15% of students from Canton Sarajevo progressed successfully from the first study year to the second study year (P12). Similarly, 20.92% of Canton Sarajevo students, among the total first-year enrollees, advanced to the third study year (S13) after successfully completing both the first and second study years. From Table 8, it can be seen that 37.27% of students from Zenica – Doboj Canton successfully completed their first study year and continued to the second study year (P12). Furthermore, 27.27% of Zenica – Doboj Canton students, who initially enrolled in the first study year, subsequently entered the third study year (S13) upon successfully completing both the first and second study years.

Table 8. Percentage of students per region who successfully completed the first and the second study year relative to the first-time enrolled students in the first study year

	CS	ZDC	CBC	PBC	TC	HNC	USC	PC	Other
P12	30.15%	37.27%	32.43%	59.09%	17.24%	68.75%	30.77%	0.00%	33.33%
S13	20.92%	27.27%	21.62%	36.36%	13.79%	50.00%	23.08%	0.00%	33.33%

Table 9 shows the percentage of students per region who successfully completed the second study year relative to students who successfully completed the first study year (P23). From Table 9 it can be seen that 69.38% of students from Canton Sarajevo and 73.17% of students from Zenica – Doboj Canton progressed successfully from the second into the third study year. It can be noted that the percentage of students from other regions is 100%. This is because there was only one student enrolled in the second study year from other regions and that student successfully finished the second study year. With regard to Posavina Canton, none of the students progressed from the first to the second year, so P23 for Posavina Canton cannot be calculated.

Table 9. Percentage of students per region who successfully completed the second study year relative to students who successfully completed the first study year

	CS	ZDC	CBC	PBC	TC	HNC	USC	PC	Other
P23	69.39%	73.17%	66.67%	61.54%	80.00%	72.73%	75.00%	NA	100.0%

3.1. Machine learning prediction of students' performance during the first study year

The relative importance of predictors (in percentages) for the performance during the first study year students is depicted in Figure 1. From Figure 1 it can be seen that the most important predictor variable is general success. Since the contribution of the most important variable is 100%, the other variables are compared to general success to determine their importance. Mathematics is 65.9% as important as general success. Physics is 65.2% as important as general success, while language is 44.2% as important as general success. Technical high school, gymnasium, and economics high school are 23.7%, 19.8%, and 16.7% as important as general success respectively. All relative importance percentages are presented in Figure 1. Other predictor variables are much less important than general success or they are with no relative importance at all as shown in Figure1.

Table 10 and Table 11 show the Confusion matrix and calculated true positive rate (sensitivity or power), false positive rate (type I error), false negative rate (type II error), and true negative rate (specificity) for the performance of the first year students. From Table 10 and Table 11, it can be seen that events and nonevents are reasonably well predicted because the true rates are relatively high and the false rates are relatively low. The true positive rate (TPR), the false positive rate (FPR), the false negative rate (FNR), and the true negative rate (TNR) on the training set are 84.0%, 19.7 %, 16.0%, and 80.3% respectively. On the test set, the same performance indicators achieved somewhat lower values with 72.7%, 22.8%, 27.3%, and 77.2% respectively. The accuracy of the model is 81.6% for the training set and 75.9% for the test set.

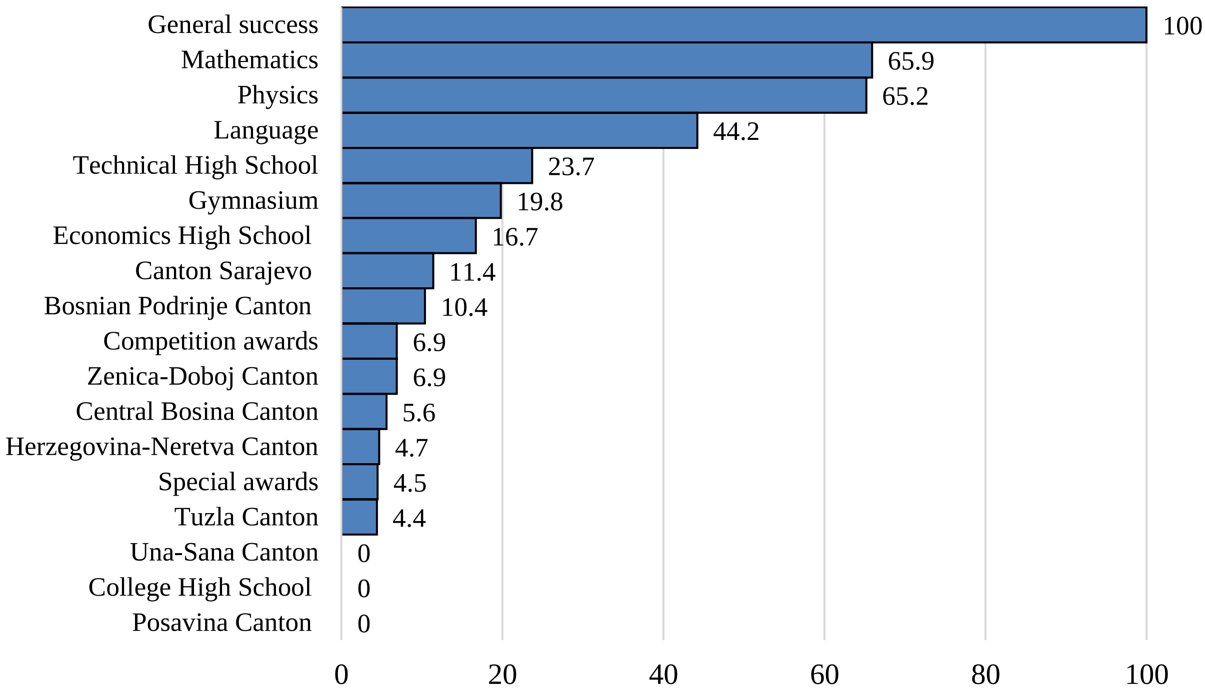

Figure 1. Relative importance of predictors (%) for the performance during the first study year

Table 10. Confusion Matrix for first study year students' performance

	Predicted Class (Training)				Predicted Class (Test)			
Actual Class	Count	1	0	%	Count	1	0	%
1 (Event)	150	126	24	84.0	33	24	9	72.7
0	295	58	237	80.3	79	18	61	77.2
All	445	184	261	81.6	112	42	70	75.9

Table 11. TPR, FPR, FNR, and TNR for first study year students' performance

Statistics	Training (%)	Test (%)
True positive rate (sensitivity)	84.0	72.7
False positive rate (type I error)	19.7	22.8
False negative rate (type II error)	16.0	27.3
True negative rate (specificity)	80.3	77.2

3.2. Machine learning prediction of students’ performance during the second study year

Figure 2 depicts the relative importance of predictors (in percentages) for the performance of the second study year students. From Figure 2 it can be seen that the most important predictor variable is the number of the courses transferred. Since the contribution of the most important variable the number of the courses transferred is 100%, the other variables are compared to the number of the courses transferred to determine their importance. GPA obtained during the first study year is 42.5% as important as courses transferred. general success, this time for the second study year students, is 16.5% as important as courses transferred. Physics and language are 14.0% % and 5.8% as important as courses transferred respectively, followed by college high school that is 3.9% as important as courses transferred.

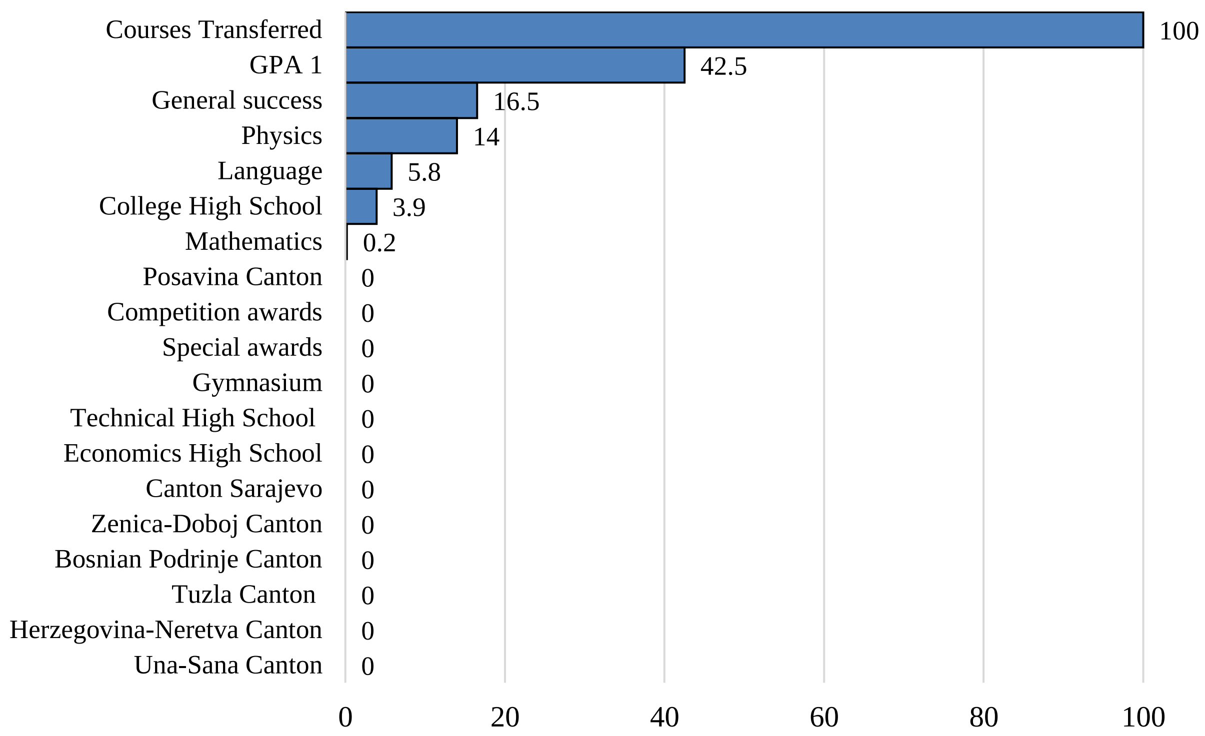

Figure 2. Relative importance of predictors (%) for the performance during the second study year

Table 12 and Table 13 show the Confusion matrix and calculated true positive rate (sensitivity or power), false positive rate (type I error), false negative rate (type II error), and true negative rate (specificity) for the performance of the second year students. From Table 12 and Table 13, it can be seen that events and nonevents are reasonably well predicted because the true rates are relatively high and the false rates are relatively low. The true positive rate (TPR), the false positive rate (FPR), the false negative rate (FNR), and the true negative rate (TNR) on the training set are 90.8%, 24.1 %, 9.2%, and 75.9% respectively. On the test set, the same performance indicators achieved somewhat lower values with 90.8%, 25.9%, 9.2%, and 74.1 % respectively. The accuracy of the model is 86.5% for the training set and 85.9% for the test set.

Table 12. Confusion Matrix for second study year students' performance

	Predicted Class (Training)				Predicted Class (Test)			
Actual Class	Count	1	0	%	Count	1	0	%
1 (Event)	131	119	12	90.8	119	12	90.8	131
0	54	13	41	75.9	14	40	74.1	54
All	185	132	53	86.5	133	52	85.9	185

Table 13. TPR, FPR, FNR, and TNR for second study year students' performance

Statistics	Training (%)	Test (%)
True positive rate (sensitivity)	90.8	90.8
False positive rate (type I error)	24.1	25.9
False negative rate (type II error)	9.2	9.2
True negative rate (specificity)	75.9	74.1

4. Conclusions

In this research machine learning prediction Decision Tree classification modeling and analysis of the students' academic performance were conducted. The analysis and measurement of students' performance were done by assessing whether students successfully completed both the first and second study years.

The most important predictor variable for the performance of the students in the first study year was general success, followed by grades in mathematics and physics, while in the second study year the most important predictor variable was the number of courses transferred from the first into the second study year. General success includes all and diverse subjects, including mathematics, physics and language, in the secondary school. These various subjects help students develop different competences and attitude which showed to be important for their academic performance. It is important to notice that general success is the third important variable when predicting the students' academic performance from the second into the third study year. Mathematics, which was the second most important variable for predicting academic performance from the first into the second study year, lost almost all of its importance. GPA in the first study showed to be very important in predicting the academic performance of the students in the second study year and can be similarly interpreted as general success in the secondary school.

Almost all students who transferred two courses from the first into the second study year didn't enroll into the third study year. That is why the number of transferred courses from the first into the second study year became the first important variable for predicting performance of the students in the second study year. This can be explained in two ways. Students have additional workload because they have more courses and because may lack the knowledge from transferred courses. Type of the secondary school and the region of the secondary school completely lost importance when predicting performance of the students from the second into the third study year. This research contributes to a better understanding of the factors that influence the academic performance of engineering students where valuable insights are provided for enrollment criteria policy decisions.

Limitation of this study was that students' performances were analyzed based on data from one faculty. Also, due to the enrollment policy change only two cohorts of students were available for analysis. Cohort of students who studied the first and the second study year during the COVID-19 pandemic were not taken into analysis because of different learning environments and different assessment methods. During the first year of the pandemic the University allowed unconditional enrollment into the next study year. Future research should focus on long term students' performances and careers, and whether secondary school and university curricula are aligned with industry and labor market demands, as well as to include include other faculties from the University.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] A. Hellas *et al.*, "Predicting academic performance: a systematic literature review," in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, Jul. 2018, pp. 175-199.

[2] S. M. Muthukrishnan, M. K. Govindasamy, and M. N. Mustapha, "Systematic mapping review on student's performance analysis using big data predictive model," *Rev. Sci. Fondam. Appl.*, vol. 9, no. 4S, p. 730, 2018.

[3] R. S. Abdulwahhab and S. S. Abdulwahab, "Integrating learning analytics to predict student performance behavior," in *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, Dec. 2017, pp. 1-6.

[4] M. M. Ashenafi, G. Riccardi, and M. Ronchetti, “Predicting students’ final exam scores from their course activities,” in *2015 IEEE Frontiers in Education Conference (FIE)*, Oct. 2015, pp. 1-9.

[5] H. Bydžovská, “A comparative analysis of techniques for predicting student performance,” in *Proceedings of the 9th International Conference on Educational Data Mining*, Jun. 2016, pp. 306-311.

[6] F. Aziz, A. W. Jusoh, and M. S. Abu, “A comparison of student academic achievement using decision trees techniques: Reflection from University Malaysia Perlis,” in *Proceedings, International Conference on Mathematics, Engineering and Industrial Applications 2014 (ICoMEIA 2014)*, May. 2015.

[7] W. Robert, J. Matthew, P. E. Miller, A. Bevlee, H. Robert, and K. Lim, “Social cognitive predictors of academic persistence and performance in engineering: Applicability across gender and race/ethnicity,” *Journal of Vocational Behavior*, vol. 94, p. 79–88, 2016.

[8] N. Kronberger and I. Horwath, “The ironic costs of performing well: Grades differentially predict male and female dropout from engineering,” *Basic Appl. Soc. Psych.*, vol. 35, no. 6, p. 534–546, 2013.

[9] S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy, “Survival analysis-based framework for early prediction of student dropouts,” in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, Oct. 2016, pp. 903-912.

[10] S. Oeda and G. Hashimoto, “Log-data clustering analysis for dropout prediction in beginner programming classes,” *Procedia Comput. Sci.*, vol. 112, p. 614–621, 2017.

[11] Y. Min, G. Zhang, R. A. Long, T. J. Anderson, and M. W. Ohland, “Nonparametric survival analysis of the loss rate of undergraduate engineering students,” *J. Eng. Educ.*, vol. 100, no. 2, p. 349–373, 2011.

[12] M. Amirah and W. Shahiri, “A review on predicting student’s performance using data mining techniques,” *Procedia Computer Science*, vol. 72, p. 414–422, 2015.

[13] F. Ahmad, N. H. Ismail, and A. A. Aziz, “The prediction of students’ academic performance using classification data mining techniques,” *Appl. Math. Sci.*, vol. 9, p. 6415–6426, 2015.

[14] M. Koutina and K. L. Kermanidis, “Predicting postgraduate students’ performance using machine learning techniques,” in *IFIP Advances in Information and Communication Technology*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 159–168.

[15] A. Abdul, “Nor Hafieza Ismailand Fadhilah Ahmad, First Semester Computer Science Students’ Academic Performances Analysis by Using Data Mining Classification Algorithms,” in *Proceeding of the International Conference on Artificial Intelligence and Computer Science*, 2014, pp. 15–16.

[16] M. Kumar India, A. J. Singh, and D. Handa, “Literature survey on student’s performance prediction in education using data mining techniques,” in *Shimla (H.P) Pin Code: 171005*, vol. 7, Summer-Hill, 2017, pp. 40–49.

[17] 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,” *Appl. Sci. (Basel)*, vol. 12, no. 19, p. 9467, 2022.

[18] W. Chang *et al.*, “Analysis of university students’ behavior based on a fusion K-Means clustering algorithm,” *Appl. Sci. (Basel)*, vol. 10, no. 18, p. 6566, 2020.

[19] A. Almasri, R. S. Alkhawaldeh, and E. Çelebi, “Clustering-based EMT model for predicting student performance,” *Arab. J. Sci. Eng.*, vol. 45, no. 12, p. 10067–10078, 2020.

[20] S. R. Sobral and C. F. de Oliveira, “Clustering algorithm to measure student assessment accuracy: A double study,” *Big Data Cogn. Comput.*, vol. 5, no. 4, p. 81, 2021.

[21] N. Varela *et al.*, “Student performance assessment using clustering techniques,” in *Data Mining and Big Data*, Singapore: Springer Singapore, Jul. 2019, vol. 1071, pp. 179–188.

[22] G. Casalino, G. Castellano, and C. Mencar, “Incremental and adaptive fuzzy clustering for virtual learning environments data analysis,” in *2019 23rd International Conference Information Visualisation (IV)*, Jul. 2019, pp. 382-387.

[23] J. Kabathova and M. Drlik, “Towards predicting student’s dropout in university courses using different machine learning techniques,” *Appl. Sci. (Basel)*, vol. 11, no. 7, p. 3130, 2021.

[24] W. Liu, J. Wu, X. Gao, and K. Feng, "An early warning model of student achievement based on decision trees algorithm," in *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Dec. 2017, p. 217-222.

[25] L. Falát and T. Piscová, "Predicting GPA of university students with supervised regression machine learning models," *Appl. Sci. (Basel)*, vol. 12, no. 17, p. 8403, 2022.

[26] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, p. 55462–55470, 2020.

[27] A. Almarabheh, M. H. Shehata, A. Ismaeel, H. Atwa, and A. Jaradat, "Predictive validity of admission criteria in predicting academic performance of medical students: A retrospective cohort study," *Front. Med. (Lausanne)*, vol. 9, 2022.

[28] S. Boumi and A. E. Vela, "Quantifying the impact of student enrollment patterns on academic success using a Hidden Markov Model," *Appl. Sci. (Basel)*, vol. 11, no. 14, p. 6453, 2021.

[29] A. Tharwat, "Classification assessment methods," *Applied computing and informatics*, vol. 17, no. 1 pp.168-192, 2020.

Appendix A

Table 1. Categorization of the types of studies with respective references

Article title	Country	Input variables	Method	Output	Reference
Integrating learning analytics to predict student performance behavior	Oman	Student grades	Compact prediction tree	Grade prediction	4
Predicting students' final exam scores from their course activities	Italy	14 behavioral features gathered from a massive open online course	Multiple linear regression	Grade prediction on a scale of 18 - 30 transformed to a grade range of 0 - 4	5
A comparative analysis of techniques for predicting student performance	Vietnam, Thailand	Demographic data, high school GPA, university GPA, and English language skills	Decision tree, bayesian network	Binary classification fail - pass, 4 class model fail, fair, good, very good	6
A comparison of student academic achievement using decision trees techniques: Reflection from University Malaysia Perlis	Malaysia	students' cumulative grade points for the first and second semesters, entry criteria, age, and gender	Decision tree	Predicting cumulative grade point average (CGPA) at the end of study	7
Social cognitive predictors of academic persistence and performance in engineering: Applicability across gender and race/ethnicity	USA	Academic support, self-efficacy, outcome expectations, interests, satisfaction, positive affect, and intended persistence at the end of each of the first four semesters.	Longitudinal analysis	GPA and additional paramters	8
The ironic costs of performing well: Grades differentially predict male and female dropout from engineering	Middle European University	Cumulative GPA, is a measure of self-doubt, measure os social discomfort, a measure of domain importance, a measure of educational experience gap, a measure of behavioral disidentification	Logistic regression, linear regression	Relationship between GPA and drop-out	9

Survival analysis based framework for early prediction of student dropouts	Detroit, USA	GPA, percentage of passed, dropped or failed credits and credit hours attempts	Time dependent Cox and Cox proportional hazards model compared with Logistic Regression, Adaboost and Decision tree	Predicting student dropout	10
Log-data clustering analysis for dropout prediction in beginner programming classes	Japan	time series data accumulated every five minutes from the history of UNIX command inputs for each class	Dynamic time warping together with clustering methods k-means, ka-medoids, k-means++	Student dropout from clasess	11
Nonparametric survival analysis of the loss rate of undergraduate engineering students	USA	Cohort group, gender, ethnic group, SAT Math score group, and SAT Verbal score group.	Descriptive statistics, Nonparametric survival analysis	Program graduation	12
The prediction of students' academic performance using classification data mining techniques	Malaysia	Gender, race, hometown, GPA, family income, university entry mode, grades Malay Language, English, and Mathematics	Decision tree, Rule-Based and Artificial Neural Network	Student academic performance	14
Predicting postgraduate students' performance using machine learning techniques	Greece	Gender, age group ([21-25]; [26-30]; [31-35]; [36- ..]) Marital Status, Number of children, Occupation, Job associated with computers (yes; no), Bachelor, Another master (yes; no), Computer literacy (yes; no), Bachelor in informatics (yes; no)	Decision tree, K-nearest neighbors using k=1, k=3, k=5, Naïve-Bayes classifier, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Random Forest, Support Vector Machines	Most efficient machine learning technique in predicting the final grade	15
First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms	Malaysia	Gender, race, hometown, family income and university entry mode	Decision tree, Naïve Bayes and Rule Based	Influencing parameters to student academic performance	16

Clustering Analysis for Classifying Student Academic Performance in Higher Education	Malaysia	Demographic data, extracurricular activities, awards, industrial training, results during studies	k-means, BIRCH, DBSCAN	Homogeneous groups	17
Analysis of University Students' Behavior Based on a Fusion K-Means Clustering Algorithm	China	Student life habits and habits during the study - life habits are described through the variables of the regular diet, physical activity, regular rest, and normal consumption (all expressed in the number of days per month), study habits are expressed through the variables of average grade, number of absences from classes, time time spent studying and time spent reading books (expressed as the number of books read during the month)	k-means i K-CFSFDP algorithm	Homogeneous groups	18
Clustering-Based EMT Model for Predicting Student Performance	Jordan	Address, year of study, age, gender, knowledge of the English language, patriotic education, management, accounting, law, Arabic, computer skills, electrical engineering, mechatronics	Canopy Cluster combined with k-means, followed by MLP classifier	Four classes: excellent, very good, good and satisfactory	19
Clustering Algorithm to Measure Student Assessment Accuracy: A Double Study	Portugal	Demographic data, and self-assessment questionnaires after completing the project during the semester	k-means, k-prototyps	Groups of students according to self-assessment ability	20
Student Performance Assessment Using Clustering Techniques	Columbia	The average grade on three tests during the semester	Fuzzy C-means	Homogeneous groups, a measure of success	21

Incremental and adaptive fuzzy clustering for Virtual Learning Environments data analysis	Open University Learning Analytics dataset	Gender, level of education, Index of Multiple Depravation, age, number of previous attempts to pass a given module, number of courses the student attends, number of submitted assignments, average grade, number of clicks on the course page	Dynamic Incremental Semi-Supervised Fuzzy C-Means	Pass or fail	22
Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques	Slovakia	Total number of accesses to the course, total points scored on all graded tasks during the course, points scored during the partial and final test	Six different classification models: logistic regression model (Logistic regression, LR), decision tree model (DT), Naive Bayes (NB), support vector machine (SVM) , Random Forest (RF) model, Neural Network (NN) model	Prediction of dropout or continuation	23
An early warning model of student achievement based on decision trees algorithm	China	The number of blogs that the student read during the course, the number of assignments completed, the number of objections that he made during the course, the number of responses to objections, the number of resources that the student accessed, the number of posts that he made during discussions on the course forums, the number of replies to posts on the forum, the number of responses posted in the objection section	Decision Tree	Predicting success at the end of the course in the form of a binary classification of fail or pass	24

Predicting GPA of University Students with Supervised Regression Machine Learning Models	Slovakia	Demographic data (age, gender, year of study, field of study, whether they have completed their studies), a questionnaire with psychological questions, questions about study habits, sociological questions (how many brothers and sisters, number of family members, whether they study alone, how comfortable they are work in groups), I will leave the questions according to the lesson, about watching videos on YouTube as a learning aid	Multinomial Linear Regression, Decision Trees, Random Forest	Student average	25
Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems	Saudi Arabia	HSGA (average high school grades), SAAT (Scholastic Achievement Admission Test), and GAT (General Aptitude Test). SAAT and GAT represent two standardized tests that all students undergo when applying to any college or university in Saudi Arabia.	Linear regressions for investigating factor strength relationships; mean prediction with: Artificial Neural Networks (ANN), Decision Tree (DT), Support Vector (SVM) and Naive Bayes	Student average	26
Predictive validity of admission criteria in predicting academic performance of medical students: A retrospective cohort study	Gulf Cooperation Council (GCC) i.e. students from Bahrain, Saudi Arabia, Kuwait, Oman, United Arab Emirates and Qatar	HSGPA (high school grade point average), AGU-MCAT (biology, chemistry, physics, and mathematics) test points as well as English language test points	Multiple regression analysis	GPA of completed first-year exams, GPA of completed fourth-year exams, assessment of basic medical sciences after 4 years of study in the form of B.Sc scores (eng. Bachelor of Medical Science exam scores) and assessment of clinical knowledge during the final	27

				phase of study in the form of MD (eng. exam scores) points	
Quantifying the Impact of Student Enrollment Patterns on Academic Success Using a Hidden Markov Model	SAD, Florida	Demographics, student admissions information, degree level achieved, courses taken as well as FAFSA reported family income information	Hidden Markov model	The impact of students' enrollment strategy on their average GPA (three strategies: full-time student, partial enrollment, mixed strategy)	28

Intentionally blank page