Predictive Factors for Completion or Dropping Out of Professional Studies

Bojan Radišić
Faculty of Tourism and Rural Development in Požega, Josip Juraj Strossmayer University of Osijek
Vukovarska 17, 34000 Požega, Croatia
bradisic@ftrr.hr

Ivan Dunđer, Sanja Seljan
Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
{idundjer, sseljan}@ffzg.unizg.hr

Abstract*. The aim of this paper is to identify the predictive factors that influence the completion or dropping out of studies. It compares the data obtained through the ISVU system between two groups of students of three-year professional studies in economics. The first group consists of students who studied at a polytechnic, whereas the second group studied at the university. First, both groups are analysed for intergroup similarity. Data comparison is done through the prediction of successful completion of studies using machine learning. Finally, an analysis identifies the differences in the predictors that influence the completion or dropping out of studies.*

Keywords. *Machine learning, multiple linear regression, success prediction, student dropout*

1 Introduction

In the Republic of Croatia, a binary higher education system was established based on the Bologna Declaration and the Act on Higher Education and Scientific Activity (Official Gazette 119/2022).

Namely, polytechnics carry out higher education and professional activities. A polytechnic can also perform scientific or artistic activities and other activities in accordance with the law and statute. A university, however, carries out higher education, scientific or artistic, and professional activities in at least two scientific areas, or a scientific and artistic area, and in at least three scientific fields. A university can, in addition, perform scientific or artistic activities and other activities in accordance with the law and statute. Polytechnics enable enrolment in professional studies that prepare students for the labour market. On the other hand, most of the universities are focused on science, but also carry out professional studies. In the academic year 2023/2024 in the Republic of Croatia, students enrolled in 116 professional studies at universities according to data on enrolment quotas in the summer enrolment period (Agency for Science and Higher Education, AZVO).

According to this data from the Agency for Science and Higher Education for the academic year 2023/2024, the number of enrolled students at the universities is satisfactory after the summer enrolment deadline. However, almost half of the enrolment places at the polytechnics in Croatia remained empty after the end of the summer enrolment period.

According to the quality standard for evaluation in the process of reaccreditation of higher educational institutions (Official Gazette 151/2022) of the Agency for Science and Higher Education, each higher educational institution must monitor and analyse the progress of students in their studies and ensure the continuity and completion of studies according to the AZVO's standard 3.3. Quality standards for evaluation in the process of reaccreditation of higher education institutions, 2023).

The conditions for enrolment and advancement of students, as well as recognition and certification, are clear, publicly announced, and consistently applied. Based on ISVU data on students from previous years, it is possible to improve the success of current and future students, which will be discussed in this article. Timely completion of studies (graduation) is important for students, their families, employers, and state authorities within the education and employment ecosystem.

The goal of this paper is to predict the success of finishing studies and to identify the factors that influence the completion or dropping out of studies.

The paper is structured in the following way. In the introductory part, the specifics of the higher education system in Croatia are presented. The second section presents related work in the field, with a special focus on suitable machine learning algorithms. The third section, i.e. the research methodology and results section, details the research workflow and findings regarding the similarities and differences between polytechnic and university students. The fourth section lists the limitations and ideas for future research, while the fifth section concludes with recommendations for university and polytechnic management.

2 Related Work

Data mining is becoming increasingly important in the education sector because it improves the education system and facilitates student development through predictive analysis. Educational institutions collect and store vast amounts of data, including student enrolment, class attendance, and exam results.

Different approaches to data mining are currently being used in the field of education with a particular focus on handling missing data. One research examined how four missing-value data sets affect machine learning algorithm performance (Radišić et al., 2023b). Arithmetic means, median values and geometric means were used to fill in the missing values, which in all cases had a positive effect.

However, researchers are increasingly favouring artificial intelligence and machine learning techniques for extracting information from educational datasets due to their superior reliability compared to other methods (Gupta et al., 2020).

Predicting student grades and grades based on their past academic data is a widely used and beneficial application in educational data mining. It serves as a significant source of information that can be utilised in various ways to improve the country's education quality (Yousafzai et al., 2020).

For instance, labelled student history data, consisting of 29 optimal attributes, was used to train a decision tree classifier and a regression model. The genetic algorithm-based decision tree classifier and the regression model showed remarkable results. The classification accuracy for grade prediction was 96.64%, while the regression-based grade prediction system had a root mean square error (RMSE) of 5.34.

Machine learning algorithms such as random forest, support vector machine, k-nearest neighbour and logistic regression (Yağcı, 2022) can be used to validly predict student success. The authors concluded that midterm exam results are essential to predict student success accurately.

Campanilla (2024) tried to predict study completion using a naive Bayesian algorithm. The findings show that of the 272 enrolled students, 200 (73.6%) had a higher GPA and were therefore more likely to complete their college education, while 72 (26.4%) were less likely to do so. The study suggested redesigning policies to support university initiatives aimed at reducing student dropout rates.

A multiple linear regression model was used to predict student success based on data obtained from surveys conducted among engineering students (Kumar et al., 2020). The results showed that four independent variables were significant for modelling students' academic success: gender, relationship with parents, previous success in studies (performance), and grades from the first semester.

A probabilistic neural network (PNN) achieved the best accuracy for predicting factors influencing student performance and success from student data extracted from a distance learning system compared to other classification algorithms (Nazif et al., 2020). It was tested with feature selection using neighbourhood component analysis for classification, and here feature selection is done using statistical measures, measures from information theory and interclass distance.

One study investigated the performance of undergraduate computer science students using data from 2001 to 2015 to predict student graduation using linear regression (LR), random forest regressor (RFR), multilayer perceptron regressor (MLPR), logistic regression (LoR), random forest (RF) and multilayer perceptron (MLP). Random forest had the best performance and linear regression performance is comparable (Alamgir et al., 2024).

A study of 13,696 Latin American university students from the first semester of 2008 to the second semester of 2020 analysed various influences on student dropout (Gutierrez-Pachas et al., 2023). A correlation analysis showed that the number of completed semesters has in all cases a robust negative correlation with dropping out.

A recent study using machine learning identified 887 students who were prone to dropping out and performing poorly in their studies. This was done using data on students' academic history, pre-enrolment status, socioeconomic status and psychological characteristics in order to determine their progress (Jayaprakash et al., 2020). The study showed that gender, family size, parental status, education and work can negatively affect student success. The authors used naïve Bayes, bagging classifier, logit boosting, random forest and enhanced iterative random forest.

Multiple linear regression is often used to predict student performance and select predictors that affect the chosen dependent variable (e.g. study completion). A recent study analysed the motivational factors and predicted the key elements that influence the duration of studies (Radišić et al., 2023a). Factors such as age, high school performance etc., best predict study duration in a multiple linear regression model.

There are several methods for selecting predictors. In an analysis of the Student Performance Dataset extracted from the University of California, Irvine (El Aissaoui et al., 2020), researchers used several methods for selecting variables, and the multivariate adaptive regression splines (MARS) method proved to be the best.

3 Research Methodology and Results

This paper deals with determining the similarities and predictive factors of success of students of professional studies at the university and polytechnic level. This is done by comparing two datasets in order to predict the success of polytechnic and university students in

completing their studies and to identify the factors that influence the completion of their studies.

Data on students were extracted from a student information system called ISVU – Information System of Higher Education Institutions (*Informacijski sustav visokih učilišta*). The data were obtained from two educational institutions that conduct three-year professional studies in the field of economics and covered the period from 2016 to 2018. This time frame was chosen because the length of the entire study, which lasted until the beginning of 2024, could be observed for the students.

The first dataset consisted of data on polytechnic students, whereas the second one covered data on university students, both from eastern Croatia.

In the first part of this research, descriptive statistics is used to show the differences and similarities between the students of both educational institutions. Here the students were analysed who have either completed their studies, dropped out of their studies or are still studying.

In the second part of this research, three machine learning algorithms for predicting the successful completion of studies are compared: random forest (RF), naïve Bayes (NB) and probabilistic neural network (PNN). This part of the research provides predictions of the success of currently enrolled students, and could enable action to be taken accordingly, in order to increase completion of studies.

In the third part of this research, the factors that influence the completion or dropping out of studies are identified, and the predictors that mostly influence the completion of studies using the multiple linear regression method. This part is essential as it can guide an institution on what kind of students to enrol, considering the influence of individual factors that affect the success of completing studies. The following hypotheses are set in this paper:

H1: Students of professional studies at the university have a shorter duration of study compared to students at the polytechnic.

H2: It is possible to identify factors that influence the completion of professional studies.

One research question was raised:

RQ1: Is it possible to predict a student's success, i.e. completion of studies?

H1 will be tested by descriptive statistics, whereas H2 will be analysed using multiple linear regression and the stepwise predictor selection method. RQ1 will be answered with the help of machine learning models.

3.1. Descriptive Data Analysis

In the first part of this research a descriptive analysis is conducted using on students who either completed their studies, gave up their studies, or who are still studying. According to the data extracted from the institutions' ISVU, in the academic years 2016/2017, 2017/2018 and 2018/2019, 207 students were enrolled at the polytechnic and 204 students at the university, which results in almost equal amounts of data in the two datasets. All data were taken at the end of the academic year 2022/2023.

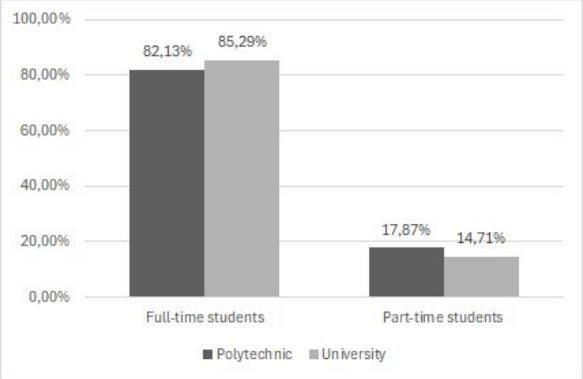

Figure 1. Type of study (enrolment type)

Fig. 1 shows the number of full-time and part-time students. The number is almost equal in both observed groups. Slightly more full-time students were enrolled at the university (85.29%), whereas slightly more part-time students were enrolled at the polytechnic (17.87%). These two groups are almost equal in terms of the number of women who studied: 62.80% at the polytechnic and 63.24% at the university. An almost equal number of students who graduated from one of the secondary economic schools enrolled at the polytechnic (54.59%) and the university (59.31%).

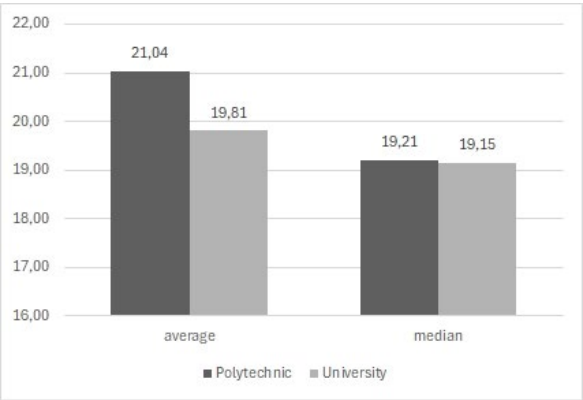

Figure 2. Age of students at study enrolment

Fig. 2 shows the average and median age at study enrolment. It can be seen that the median age is almost the same at the polytechnic (19.21) and university (19.15). There is, however, a difference in the average age at enrolment – namely, the average age at the polytechnic is slightly higher (21.04) than at university (19.81). Students older than 28 did not choose to study at the university but enrolled at the polytechnic. The oldest student who decided to enrol at the university was younger than 28 years old. For comparison, 17 students over the age of 28 were enrolled at the polytechnic, which resulted in a slightly higher average age of students who enrolled at the polytechnic.

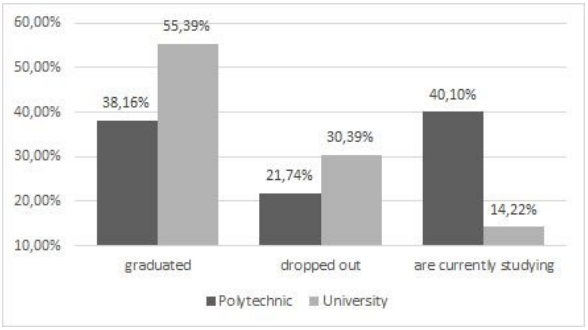

Figure 3. State of completion of professional studies

Fig. 3 shows the state of completion of professional studies among university and polytechnic students. Out of the total number of students enrolled, more than half (55.39%) completed their studies at the university, while slightly more than a third (38.16%) completed their studies at the polytechnic. Nevertheless, dropping out of studies is more pronounced at the university (30.39%), where almost every third of students drops out, while at the polytechnic almost every fifth student drops out (21.47%).

The most significant difference between these two groups is the number of students who are still studying. It is 40.10% at the polytechnic and 14.22% at the university, which affects the duration of studies and thus increases the projected (expected) duration of studies, which does not favour the polytechnic in this research.

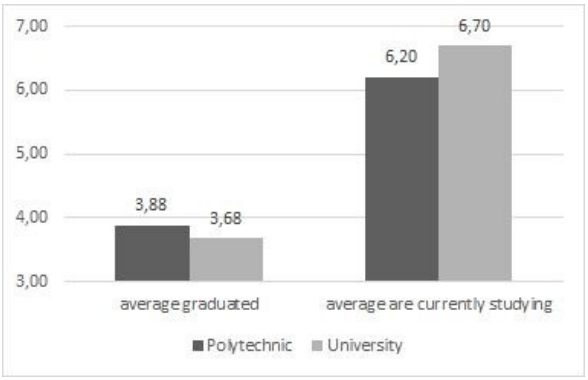

Figure 4. Average duration of study

Fig. 4 shows the average duration of studies, but with a note that less than 60% of the polytechnic students have completed their studies, while the rest have either dropped out or are still studying.

That part is more representative for the university and refers to slightly less than 86% of the total sample who completed their studies. The average duration of studies of students who have completed their studies (which officially lasts 3 years) is 3.88 years for the polytechnic and 3.86 years for the university. Students who are still studying study for an average of 6.20 years at the polytechnic and 6.70 years at the university, which is almost the same average.

If the weighted average of those who have graduated and those who are still studying is calculated, assuming that all those who are currently studying complete their studies, the average duration of studies is 4.81 years at the polytechnic and 4.1 years at the university. According to the weighted average, university students complete their studies in 15% less time than polytechnic students. This confirms the hypothesis (H1) that students of professional studies at the university have a shorter duration of study compared to students at the polytechnic.

Fig. 5 shows the most significant difference between students who choose the university over the polytechnic. The overall results of the state exam (state graduation/matriculation) show that the students who chose to study at the university achieved an average of 341.16 points, which is 52.34% more than the average of 223.95 points, achieved by students who chose to continue their education at the polytechnic. It should be noted here that 44 students entered the polytechnic without taking the state exam.

Furthermore, there are many similarities in the characteristics of enrolled students at the university and the polytechnic. An equal number of full-time students are enrolled; they come from the same county where they study and have approximately the same average age, and there are equally many who have graduated from one of the secondary schools of economics.

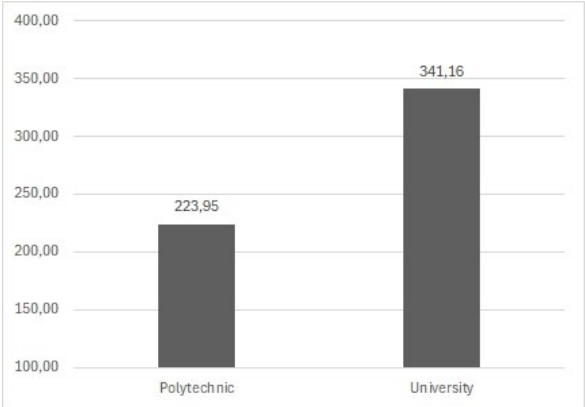

Figure 5. Average state exam score

On the other hand, the most significant difference is in the results they achieved on the state exam, where the students who enrolled their studies at the university achieved significantly better results. Students who had better success at the state exam enrolled at the university. This resulted in a shorter duration of studies and faster entry into the labour market or continuation of education.

3.2. Machine Learning

In the second part of this research, three machine learning algorithms were compared to determine their applicability in predicting student success in studies. The output variable was the completion of studies divided into three categories:

1) graduated students (**GS**),
2) students who dropped out from studies (**DO**),

3) students who are currently studying (**AS**).

Nine input variables (i.e. data features) were selected and are listed and described in Table 1.

Table 1. Data features

Features	Description
Enrolment type	Part-time students, Full-time students
Age	Student's age at the time of study enrolment
Gender	Male, Female
County	County of residence
County Yes/No	Residence and place of study is in the same county or not
Secondary school	Type of finished secondary school
Economics Yes/No	Completed secondary school of economics or not
Score	Secondary school final score
Exam score	State exam score

Machine learning (ML) methods can be used to predict study completion or dropout. Namely, the most commonly used methods include random forest, naïve Bayes, and neural networks, which are explored in this study.

Random Forest (RF)

The random forest (RF) method is usually employed for classification and regression tasks (Kovač et al., 2022). It is a straightforward and reliable classifier that exhibits a number of beneficial attributes. During the training phase, the algorithm generates several decision trees using bootstrapping (sampling with replacement) and random feature selection. When dealing with a new data point, each decision tree makes a prediction. In the case of classification, the final output is determined by taking the mode (the most frequently predicted class). In regression, the final result is the average of all the predictions. It provides advantages in automating the handling of missing data values and efficient processing of large datasets. Certain disadvantages are associated with increased computing requirements and the use of resources needed to achieve optimal results.

Naïve Bayes (NB)

Naïve Bayes (NB) is a classification algorithm based on Bayes' theorem (Berrar, 2019):

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

P(A) is the probability of A occurring, P(B) is the probability of B occurring, P(A|B) is the probability of A given B and P(B|A) is the probability of B given A. The naïve Bayes algorithm is very versatile and practical, making it generally preferred for categorising large datasets. In certain cases, the feature independence assumption occasionally needs to be corrected, leading to suboptimal classification results. The classifier is probabilistic, which means that it makes predictions based on the probability of an object occurring.

Probabilistic Neural Network (PNN)

Probabilistic neural networks (PNNs) are artificial neural networks created using the Parzen approach (Mohebali et al., 2020). The purpose of PNNs is to develop a collection of estimators for the probability density function. These estimators are designed to approximate the Bayesian optimal minimisation of "expected risk", popularly known as "Bayesian strategies". Probabilistic neural networks (PNNs) have demonstrated significant potential in solving complex scientific and technical problems.

Results

Both student datasets (containing data on polytechnic and university students, respectively) were compared using all three machine learning algorithms to compare their accuracy. 80% of the data was randomly left for training, and the remaining 20% was used for testing the algorithms.

Results for polytechnic students are presented in Fig. 6, which shows that the algorithms themselves did not achieve high accuracy. The NB and RF algorithm were almost equal, with an accuracy of 45.20% for NB and 42.86% for RF, while the PNN showed a lower accuracy of 33.33%.

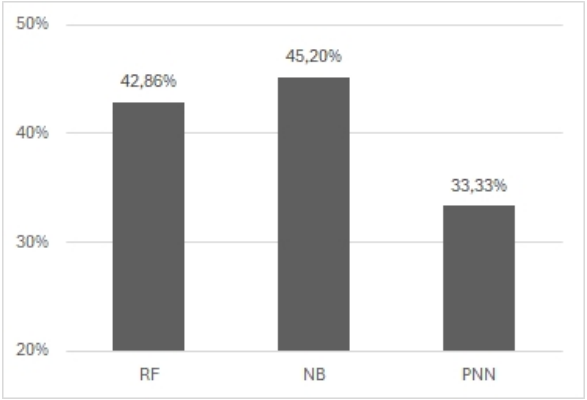

Figure 6. Accuracy for polytechnic students

Although the results for polytechnic students are quite low, a closer examination of the confusion matrices in Table 2 shows that NB exhibited a high accuracy of 83.30% in recognising students who are still studying. RF (66.70%) and PNN (60%) also recognised students who are still studying better than graduates or dropouts, but it was lower than for NB.

All three machine learning algorithms were least able to recognise students who dropped out of their studies; RF recognised every fifth, NB every fourth, and PNN recognised none.

Fig. 7 shows the accuracy results of the machine learning algorithms for university students. The most accurate was NB, with 63.42%, followed by less accurate RF algorithm (58.54%). Both had better accuracy compared to the same algorithms when used on the polytechnic students' dataset. As in the former case, the PNN algorithm was the least accurate, with 48.78%.

Table 2. Confusion matrix for polytechnic students

	True data/Predicted data	GS	DO	AS
RF	**GS**	5 (26.31%)	4	5
	DO	4	0 (00.00%)	2
	AS	10	4	9 (56.25%)
NB	**True data/Predicted data**	**GS**	**DO**	**AS**
	GS	12 (46.15%)	0	1
	DO	4	2 (20.00%)	0
	AS	10	8	5 (83.33%)
PNN	**True data/Predicted data**	**GS**	**DO**	**AS**
	GS	5 (26.31%)	4	4
	DO	4	0 (00.00%)	2
	AS	10	4	9 (60.00%)

Remarks: graduated students (**GS**), students who dropped out from studies (**DO**), students who are currently studying (**AS**)

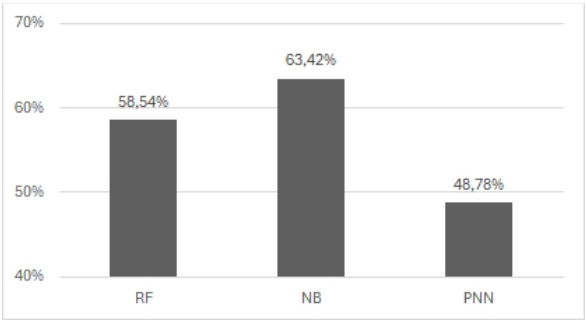

Figure 7. Accuracy for university students

Table 3 shows the confusion matrix of all machine learning algorithms for university students. The NB algorithm showed the highest accuracy of 76.90% in recognising university students who graduated. The RF and PNN algorithms showed almost similar accuracy results in recognising students who graduated, with 65.70% for RF and 62.50% for PNN. All three machine learning algorithms were least able to recognise students who are still studying; RF and PNN did not recognise any of the cases, whereas NB recognised 50%.

The results of this subsection can help educational institutions to identify the possible duration of studies of currently enrolled students. The use of the NB algorithm is particularly applicable and useful for the polytechnic, which achieved an accuracy of (83.30%) in recognising students who are still studying at the polytechnic.

The algorithms have the lowest accuracy when identifying students who dropped out of their studies, the NB algorithm has an accuracy of 20%, while the RF and PNN failed to recognize a single student of the polytechnic.

Table 3. Confusion matrix for university students

	True data/Predicted data	GS	DO	AS
RF	**GS**	22 (65.70%)	1	1
	DO	9	1 (20.00%)	0
	AS	3	3	0 (00.00%)
NB	**True data/Predicted data**	**GS**	**DO**	**AS**
	GS	20 (76.90%)	4	1
	DO	5	5 (38.50%)	0
	AS	1	4	1 (50.00%)
PNN	**True data/Predicted data**	**GS**	**DO**	**AS**
	GS	15 (62,50%)	9	1
	DO	5	5 (31.25%)	0
	AS	4	2	0 (00.00%)

Remarks: graduated students (**GS**), students who dropped out from studies (**DO**), students who are currently studying (**AS**)

By timely detection of such students, it is possible to take some measures that could help students shorten their duration of study and thereby significantly shorten the overall average duration of studies at their institution. In this case, the university could use the NB algorithm (accuracy shown to be 76.90%) to identify students expected to graduate successfully. This can also identify those students who will either drop out or continue studying and focus on them to help them finish their studies.

The algorithms have the lowest accuracy in the case of identifying dropped out students, the NB algorithm has an accuracy of 38.50%, while the RF (20.00%) and PNN (31.25%) have lower accuracy for university students. This answers RQ1, since it possible to predict to a certain extent the students' success in completing their studies (e.g. 76.90% for university students with the help of NB).

3.3. Multiple Linear Regression

In this subsection, the authors determine what affects the completion of studies. Ten variables were selected from the ISVU database from each of the two selected institutions: completed studies (Yes/No), enrolment type (part-time or full-time students), age (student's age at the time of study enrolment), gender (male/female), county of residence, residence and place of study is in the same county (Yes/No), type of finished secondary school, completed secondary school of economics (Yes/No), secondary school final score and state exam score.

This third part of the research investigated the correlation between the completion of studies (i.e. the dependent variable), and the other mentioned

variables, also known as predictors or predictor variables.

Multiple linear regression (MLR) is a statistical technique for creating a model that represents the relationship between two or more independent and dependent variables by fitting a linear equation to observed data. The underlying theoretical assumption of MLR is that any unit change in the independent variable results in a consistent change in the dependent variable (El Aissaoui et al., 2020). The multiple linear regression equation is shown below:

$$Y = k_0 + k_1x_1 + k_2x_2 + \cdots + k_nx_n$$

Y=dependent variable,
x_i=predictors,
k_i=slope coefficient for every predictor,
k_0=intercept (the value of Y when all predictors are zero).

Table 4 shows a comparison of two multiple linear regression models for the polytechnic students using the stepwise method. The R-squared value of the second model is 0.099, indicating that the model explains 9.99% of the variation in graduation (i.e. the completion of studies). In the second model, the predictors are enrolment type (part-time or full-time students) and gender, and form the most efficient multiple linear regression model between the dependent variable (completion of studies) and these two mentioned predictors.

Table 4. Comparison of models for polytechnic students

Model	R	R²	R² Change	F Change	df1	df2	p
1	0.256	0.066	0.066	14.424	1	205	< .001
2	0.315	0.099	0.034	7.616	1	204	0.006

1. Predictors: (Constant), Enrolment type
2. Predictors: (Constant), Enrolment type, Gender

In Table 5, the column "p" shows that both predictors make a statistically significant unique contribution (<0.05) to the final value (case of polytechnic students).

Table 5. Correlation between the completion of studies and other predictors for polytechnic students

Coefficients						
Model		Unstandardized	Standard Error	Standardized	t	p
2	(Constant)	1.368	0.325		4.211	< .001
	Enrolment type	0.595	0.152	0.260	3.913	< .001
	Gender	-0.333	0.121	-0.183	-2.760	0.006

The highest absolute value is scored by enrolment type (part-time or full-time students) which has the greatest influence of 0.260. This means that if the predictor reflecting enrolment type (part-time or full-time students) increases by one standard deviation, and other predictors remain constant, then completion of studies will increase by 0.260 of standard deviations. Also, the influence of gender has negative sign and amounts to -0.183. This means that if the predictor that reflects gender increases by one standard deviation, and other predictors remain constant, then completion of studies will decrease by 0.183 standard deviations. The model implies that full-time female students will eventually graduate.

Table 6 shows a comparison of three multiple linear regression models for the university students using the stepwise method. The R-squared value of the third model is 0.230 indicating that the model explains 23% of the variation in graduation (i.e. the completion of studies). In the third model, the predictors are gender, enrolment type (part-time or full-time students) and type of finished secondary school, and form the most efficient multiple linear regression model between the dependent variable (completion of studies) and these three predictors.

Table 6. Comparison of models for university students

Model	R	R²	R² Change	F Change	df1	df2	p
1	0.335	0.112	0.112	25.468	1	202	< .001
2	0.433	0.188	0.076	18.797	1	201	< .001
3	0.480	0.230	0.042	11.045	1	200	0.001

1. Predictors: (Constant), Gender
2. Predictors: (Constant), Gender, Enrolment type
3. Predictors: (Constant), Gender, Enrolment type, Type of finished secondary school

In Table 7 the column "p" shows that all three predictors make a statistically significant unique contribution (<0.05) to the final value (case of university students).

Table 7. Correlation between the completion of studies and other predictors for university students

Coefficients						
Model		Unstandardized	Standard Error	Standardized	t	p
3	(Constant)	2.105	0.288		7.305	< .001
	Gender	-0.434	0.094	-0.288	-4.68	< .001
	Enrolment type	0.617	0.129	0.301	4.792	< .001
	Type of finished secondary school	-0.076	0.023	-0.208	-3.323	0.001

The highest absolute value is scored by enrolment type (part-time or full-time students) which has the highest influence of 0.301. This means that if the predictor that reflects enrolment type (part-time or full-time students) increases by one standard deviation, and other predictors remain constant, then completion of studies will increase by 0.301 of standard deviations. Also, the influence of gender and type of finished secondary school have negative signs; -0.288 for gender and -0.208 for the type of finished secondary

school. This means that if the predictor that reflects gender increases by one standard deviation, and other predictors remain constant, then completion of studies will decrease by 0.288 of standard deviations and decrease by 0.208 for the type of finished secondary school. The model implies that full-time female students that finished a secondary school of economics usually graduate.

In this part, it can be concluded that gender and enrolment type (part-time or full-time students) affect the success of completing studies both at the polytechnic and the university. However, for the institutions themselves, these predictors have very little significance for attracting future students, except for increasing the enrolment quotas for full-time students.

Nevertheless, the financial benefits made by the educational institutions would make it unfavourable to reduce the number of part-time students. Another important aspect is that studying part-time suits older students because, in addition to work, it enables them to continue their education and training to acquire new knowledge that they will use in their workplace.

More significant for the university is that more successful students finished a secondary school of economics, i.e. they will be more successful and eventually graduate. The university can therefore adjust and strengthen its marketing activities there to attract as many students as possible.

Factors that significantly influence the completion of studies were determined by multiple linear regression. For polytechnic students, these are type of enrolment and gender, and for university students: gender, type of enrolment, and type of completed high school, thereby confirming H2.

4 Limitations and Future Research

This research has limitations regarding the number of input variables as a more significant number would allow more accurate predictions with the help of machine learning algorithms. Data on student success by semester and academic year could be included as well, and it would be beneficial to include other elements of monitoring students during their studies (attendance of classes, homework outcomes, colloquia result, academic engagement metrics etc.). For future research, it would be advisable to include a longer time period (i.e. longitudinal data and across more cohorts) in order to improve the robustness of the results with regard to temporal variations.

In addition, it would be wise to include data from more institutions from the same field for future research to obtain more representative data. It would also be valuable to compare and include educational institutions from different scientific fields. This would help to better generalize the findings and validate the models across diverse educational settings.

5 Conclusion and recommendations

Every higher educational institution in Croatia should monitor and analyse the progress of students, and ensure the continuity and completion of their studies. This paper showed that it is possible to predict students who will be successful in their studies and who will be able to graduate. The duration of the full study, which continued until the start of 2024, could be seen for the students, therefore that is why this time frame was selected.

The second part of this research showed that institutions can detect students who are currently enrolled and actively studying. As shown in this paper, students who are still studying and who, on average, will potentially study longer can be recognized at the polytechnic.

The polytechnic can enable such students to study for a shorter period time, possibly through various forms of assistance in learning and teaching. The university, on the other hand, can identify students who will complete their studies. However, those remaining students can be provided with different forms of assistance in learning and teaching in order to help them finish their studies more successfully.

This research has shown that students of professional studies at the university level have a shorter study time (4.1) than students at the polytechnic (4.81).

The research identified the following predictors as factors that influence the completion of studies: enrolment type (part-time or full-time students) and gender at the polytechnic; and at the university: gender, enrolment type and type of finished secondary school.

Using the naïve Bayes machine learning model, it is possible to identify polytechnic students who are still studying with an accuracy of 83.30% and with an accuracy of 76.90% the university students expected to graduate successfully.

Based on the data that higher educational institutions collect and store about their students, they can direct their marketing activities and the promotion of their institution to a precisely defined population of potential future students. Activities should be focused on students that finished a secondary school in the field of economics to attract students who are successful in their studies.

Finally, the recommendations to the polytechnic go in the direction of using machine learning to detect students who are expected to study longer and to take measures related to improving the academic efficiency of the identified students. Using machine learning, the university can identify students who will successfully complete their studies and, just like polytechnics, can take measures to improve the academic performance of other students. Furthermore, the university should work on attracting a larger number of students who completed their high school education in the field of economics.

References

Alamgir, Z., Akram, H., Karim, S., Wali, A. (2024). Enhancing Student Performance Prediction via Educational Data Mining on Academic Data. *Informatics in Education*, 23(1), 1–24, doi:10.15388/infedu.2024.04

Agency for Science and Higher Education (AZVO) (Agencija za znanost i visoko obrazovanje). Retrieved from https://www.studij.hr/broj-prijava

Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier. *Encyclopedia of Bioinformatics and Computational Biology*, 1, 403–412. doi:10.1016/B978-0-12-809633-8.20473-1

Campanilla, B. S. (2024). Forecasting Degree Completion: A Naïve Bayes Predictive Model for Students' Success. In *Proceedings of the Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1–4). doi: 10.1109/ICAECT60202.2024.10469398

El Aissaoui, O., El Madani, Y. E., Oughdir, L., Dakkak, A. & El Allioui, Y. (2020). A Multiple Linear Regression-Based Approach to Predict Student Performance. In M. Ezziyyani (Ed.), Proceedings of the *Advanced Intelligent Systems for Sustainable Development Conference (AI2SD'2019)* (pp. 9–23), Advances in Intelligent Systems and Computing, vol. 1102. Springer, Cham. doi:10.1007/978-3-030-36653-7_2

Gupta, S. B, Yadav, R. K. & Gupta, S. (2020). Analysis of Popular Techniques Used in Educational Data Mining. *International Journal Of Next-Generation Computing*, 11(2), 137–162. doi:10.47164/ijngc.v11i2.178

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G. & Gomez-Nieto E. (2023). Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context. *Education Sciences*, 13(2), 154. doi:10.3390/educsci13020154

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting Students Academic Performance using an Improved Random Forest Classifier. In *Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 238–243). doi:10.1109/ESCI48226.2020.9167547

Kovač, A., Dunđer, I. & Seljan, S. (2022). An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services. In *Proceedings of the 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 954–961). doi:10.23919/MIPRO55190.2022.9803517

Kumar, A., Eldhose, K. K., Sridharan R. & Panicker, V. V. (2020). Students' Academic Performance Prediction using Regression: A Case Study. In *Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1–6). doi:10.1109/ICSCAN49426.2020.9262346

Mohebali, B., Tahmassebi, A., Meyer-Baese, A. & Gandomi, A. H. (2020). Probabilistic neural networks: a brief overview of theory, implementation, and application. In: P. Samui, S. Chakraborty, D. T. Bui, R. C. Deo (Eds.), *Handbook of Probabilistic Models*, Elsevier, Amsterdam (pp. 347–367). doi:10.1016/b978-0-12-816514-0.00014-x

Nazif, A. M., Hesham Sedky, A. A. & Badawy, O. M. (2020). MOOC's Student Results Classification by Comparing PNN and other Classifiers with Features Selection. In *Proceedings of the 21st International Arab Conference on Information Technology (ACIT)* (pp. 1–9). doi:10.1109/ACIT50332.2020.9300123

Official Gazette 119/2022 (NN 119/2022, Zakon o visokom obrazovanju i znanstvenoj djelatnosti). Retrieved from https://narodne-novine.nn.hr/clanci/sluzbeni/full/2022_10_119_1834.html

Official Gazette 151/2022 (NN 119/2022, Zakon o osiguravanju kvalitete u visokom obrazovanju i znanosti). Retrieved from https://narodne-novine.nn.hr/clanci/sluzbeni/2022_12_151_2330.html.

Quality standards for evaluation in the process of reaccreditation of higher education institutions (Standardi kvalitete za vrednovanje u postupku reakreditacije visokih učilišta). Retrieved from https://www.azvo.hr/wp-azvo-files/uploads/radne-grupe/12/20/Standardi-kvalitete-za-vrednovanje-u-postupku-reakreditacije-visokih-ucilista.pdf

Radišić, B., Dunđer, I. & Seljan, S. (2023a). Data Analysis of the Motivation and Factors for a Shorter Duration of Study. In *European Conference on Information and Intelligent Systems (CECIIS)* (pp. 179–187). Varaždin: University of Zagreb, Faculty of Organization and Informatics Varaždin.

Radišić, B., Seljan, S. & Dunđer, I. (2023b). Impact of missing values on the performance of machine learning algorithms. In *CEUR Workshop Proceedings: Recent Trends and Applications in Computer Science and Information Technology (RTA-CSIT 2023)* (pp. 54-62). Tirana: University

of Tirana, Faculty of Natural Sciences, Department of Informatics.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environment*, 9(11). doi:10.1186/s40561-022-00192-z

Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Education and Information Technologies, 25, 4677–4697. doi:10.1007/s10639-020-10189-1