The 22nd IAMUC
Batumi, 21-22 Oct 2022

Proceedings of the
International Association of Maritime Universities Conference

Identification of Features Associated with University Dropout-a case study of University of Split, Faculty of Maritime Studies

Anita Gudelj[1,*], Helena Ukić Boljat [1] and Merica Slišković [1]

[1] *IAMU Faculty of Maritime Studies, University of Split, 21000 Split, Croatia*
[*] *Corresponding author: anita@pfst.hr; Tel.:* +385-21-619-475.

Abstract: The primary goal of higher education institutions is to provide a quality educational process. Maritime higher education is an essential element in acquiring the knowledge, and skills needed on board a ship. One of the indicators of potential problems in this educational process may be a high number of dropouts in the early years. Predicting student success and dropout, or identifying students who are at higher risk for dropping out, is critical to improving the quality of higher education. An analysis of the academic performance of students at the University of Split, Faculty of Maritime Studies (PFST) was conducted a high dropout rate was revealed. This research aims to improve early prediction of student dropout by identifying the most relevant features. The data is processed and the features that influence dropout are extracted through an attribute selection algorithm and machine learning techniques such as a random forest. The results of our research suggest that higher education institutions should be aware of the need to identify early the profile of students who are at potential risk of dropping out. Moreover, the developed model is useful for strategic planning of additional mechanisms to improve the efficiency of study at maritime higher education institutions.

Keywords: maritime higher education; dropout; machine learning; feature selection; Faculty of Maritime Studies Split

1. Introduction

Term dropout refers to students leaving their university studies before having completed their study program and obtained a degree. [1] Dropout can be considered as one of the key indicators of the quality and attractiveness of the educational program, which emphasizes the existence of serious difficulties and barriers in the high educational system that students face to graduate. It could be a serious problem for the students, their families, high educational institutions, labor market and the society in generally.

According to the latest data from countries that belong to the Organization for Economic Cooperation and Development (OECD), many young people leave education between the ages of 18 and 24. The latest data shows that 47% (almost half of them) have left education system. The average age of first time tertiary graduation was 25, while 86% of them graduated before they turned 30. In addition, women are more likely to enter the tertiary education between ages of 25 than man is. The most common tertiary qualification is Bachelor degree. [2] In 2019, the OECD highlighted that on average 12% of students who entered a full-time bachelor's program, have left the tertiary system before beginning their second year of study. Furthermore, this share increases to 20% by the end of the program's theoretical duration and to 24% three years later. [3] Percent of students who dropped out of the higher education system vary across the countries and regions. For example, authors in the report from Latin America and Caribbean region state, despite the concentration of dropouts at the beginning of their student careers, almost 30 percent of all dropouts leave the system after spending four years in it. [4]

For this reason, many universities within European Space of Higher Education take into account in their strategic plans reduction of the rate of students' dropout as main objectives. [5] The national quality assurance agency in its requirements specified in European Qualification Framework (EQF) considers dropout rate as one

of measurable factor of studying success. For this reason, the University of Split and Faculty of Maritime Studies, as one of its faculties, take into account in their strategic plans as main goals to reduce the rate of students' dropout.

The students that enroll Faculty of Maritime Studies Split have various educational background (various high schools and education profiles, different levels of success measured according to average grades at schools or state examinations). As stated in [6] combination of these factors, together with current student engagement probably have effect on their success in the early phase of their studies. Predicting the student's success in the early phase of their studies helps faculties in redirecting activities to less performing students in order improve their success. Similar is with dropout rate. It is of genuine importance to limit dropout, and therefore, the aptitude to predict students' dropping out could be very useful. [5]

Dropout can be caused by different factors. Some of them can be the family and personal reasons, poor level of previous technical knowledge that depends on graduated secondary school, poor academic performance, and low motivation rate. As is stated in [7], the higher risk of abandoning studies is implied for students with weak educational strategies and without perseverance to attain their aims in life. These students also have low academic performance and low success rates.

This paper aims to investigate the potential benefit of using the feature selection algorithm for enhancing the classification accuracy of the applied classifier to identify at risk students in advance and help them. The research focuses on multi-classification and developing machine learning predictive models for diagnosing student dropout. The objective is to identify the profile of students who tend to drop out. Three classes represent the students: ACTIVE, GRADUATED, and DROPPED.

2. Research approach and methodology

The methodology proposed in this study for predicting the features affecting the dropout of students belongs to the process of Machine Learning. 70% of the data was used to test the accuracy of the model, and 30% of it was used as training data. The workflow of research approach is shown in Figure 1.

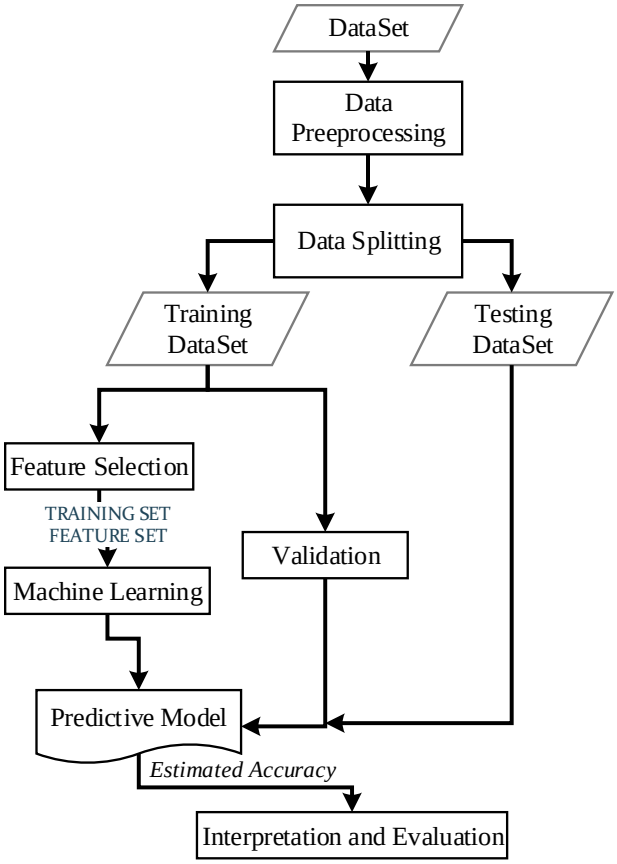

Figure 1. Research workflow

2.1 Data Collection

The data were gained from the student's database available in the Information Sys-tem of Higher Education Institutions (ISVU) introduced in the University of Split, Faculty of Maritime Studies. The data were taken for the pure generations comprising only students who had been enrolled to the first semester (first academic year) for the first time. Due to the fact that there are no registered dropped students at the Master's level, the analyzed data specifically covered the Bachelor's level in five study programs: Marine Engineering (BS), Nautical Studies (PN), Marine Electrical Engineering and Information Technologies (PEIT), Maritime Yacht and Marina Technologies (PTJM), and Maritime Management (PM). Since, the expected time of graduation of students is between 3 and 4 years (in the faculty, each study program under analysis has the curricula, with six semesters), from a temporal point of view, the data were extracted for six academic years (from 2012 to 2018). In the considered period, it has been observed that student's study above expected time of graduation (for 4 years and more). Using this time framework, 1436 student's records have been collected, 1095 males, and 341 females, according to Figure 2. The most students studied Nautical study, and the most of them graduated gymnasiums.

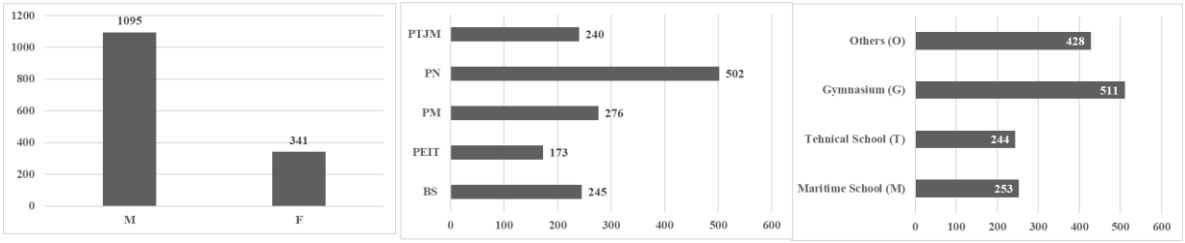

Figure 2. Gender, study and graduated school histogram graphs

2.2 Data preprocessing and splitting

The initial dataset consisted of 1436 records and 30 features. Data preprocessing included following actions: cleaning features, removing redundant features, handling improperly formatted and data normalization. Some of features associated with student's personal information are cleaned. Additionally, some features were irrelevant for our study e.g., the "Post Code", "Birth place", "Faculty name" and "Faculty identification number" and they have been also removed from started dataset. To identify and remove redundant features, all features were transformed into numerical and the pairwise correlation coefficients were calculated evaluating associations among features as shown in Figure 3.

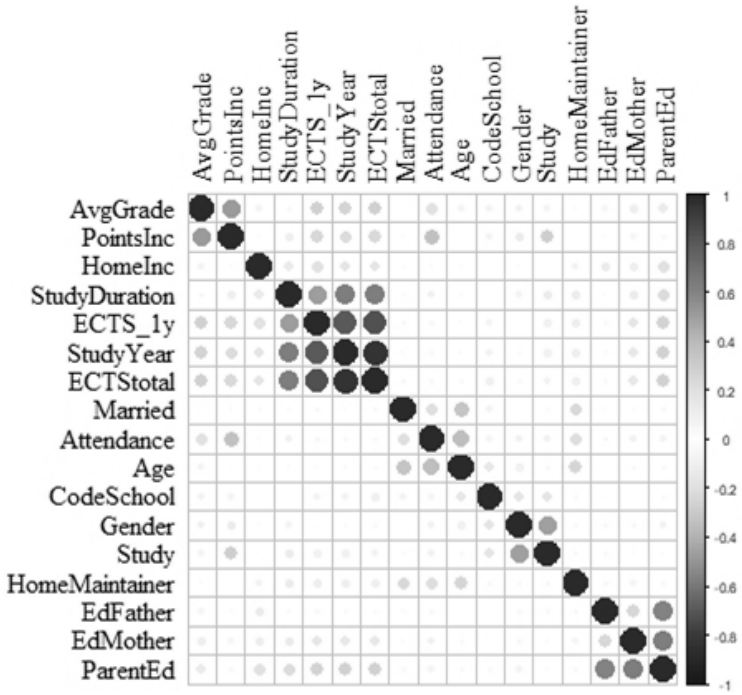

Figure 3. Correlation coefficients for all features

The color intensity is proportional to the correlation coefficient. Thus, the stronger the correlation (i.e., the closer -1 or 1), the darker the square. Two pairs of features showed a high correlation coefficient: “StudyYear” and “ECTStotal”, “ECTStotal” and “ECTS_1y”. “StudyYear” and “ECTStotal” were subsequently removed since their values were highly correlated with the feature “ECTS_1y” (> 0.7).

After data pre-processing phase, the final dataset contains 18 features. The feature "DropOut" is the target variable and 17 remaining features are the predictors. These features are summarized in the Table 1.

Table 1. Overview of features with the description

Feature	**Metadata**
Gender	Sex of the students
Age	The age when a student enrolled the study
Married	Marital status of the student
HomeMaintainer	It includes information about who maintains the student. Also it contains information of family support
EdMother, EdFather ParentEdBackground	Educational background of student's mother, father and parents/guardian, respectively
HomeInc	A description of the occupational position of the father and mother
CodeSchool	It refers to type of a high school which students graduated
AverageGrade	It refers to the grade average to two decimal places in four grades of high education
PointsInc	It represents the total points achieved at the end of the high education program. It includes points gained from high school and the final state exam
DropOut	A categorical variable which denotes whether the student graduated, still studies or dropped
Study	Type of graduated secondary school
Attendance	Mode of study
StudyYear	The last academic year in which the student enrolled
StudyDuration	Estimated values from first year of admission and last year of admission
ECTS_1y	The number of credits that the student gained during the first study year (the first enrolment).

Since the purpose of the research is to predict the regularity of dropouts and retentions, three categories of students were identified and therefore the target label "DropOut" has three categories.

Active: This category involves students who enrolled in the last academic year and have not yet graduated. A total of 422 (or 29.39% of all instances) students belonged to this category.

Graduated: involves students who successfully finished their study. 569 (39.62 %) observed students finished her study.

Dropped students: This class involves students who or were registered as dropped on personal request or have had at least two academic years without enrolling and they have not yet graduated. 445 (30.99 %) students were considered as dropped. The analysis revealed that there is no scientific difference between the percentage of female students and male students who dropped out (~ 30%). 44% of females and 38% of males graduated.

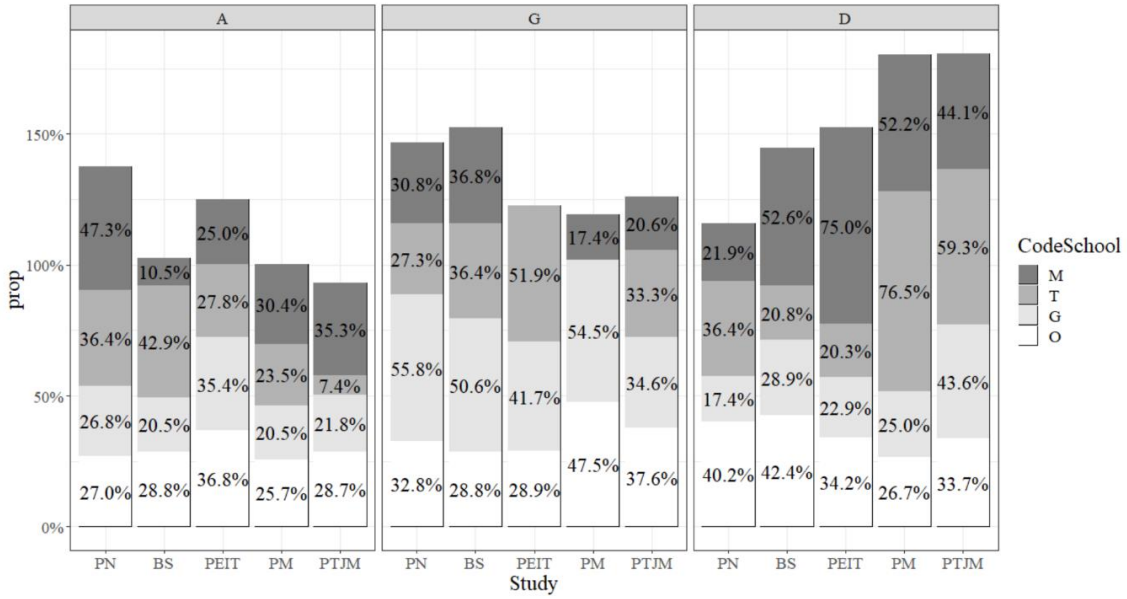

Figure 4. Student status according to their study and graduated school: A-active students, G-graduated students, D-dropped students

2.3. Feature Selection

Machine learning techniques use features (variables) to generate predictive models selected by applied feature selection algorithm. Feature selection consists of identifying the relevant features for building a predictive model with the goal of obtaining a suitable combination of features that will increase the application's performance. The use of the proper feature selection algorithm reduces the size of the dataset and might have effect of increasing the precision and accuracy of developed model, reduces requirements for computer memory and gains speed. In this study three different feature selection methods were applied creating three different feature subsets. The Recursive Feature Elimination (RFE) method is a recursive feature selection approach. This method effectively selects features in a training dataset that are more or the most relevant for the prediction of target variable. The RFE method recursively ranks features regarding some metrics of their importance. At each iteration feature importance was measured. The less relevant one was removed, and model was built on the remained features. [8]

Genetic algorithm (GA) is an algorithm based on the mechanisms of natural selection and natural genetics. GA applies the law of the survival of the fittest to the in-formation sharing system and thus improving the performance of searching information. Instead of randomly collecting information, extracted genetic algorithms, by using efficiently collected information predict new search points for improvement of the performance.[9] GA are often applied in data mining where GA can be implemented as the classifier or as a result of the optimizer. [10]

The Boruta algorithm was developed to determine all important features within a classification framework and it works on random forest (RF) method.[11] Boruta tries to select a set of important features by comparing the importance of the real predictors with those of random predictors using statistical testing and several runs of RFs. Unimportant features and randomize variables are removed and the previous steps are repeated until the number of RF running don't reach set maximal number of times RF is running.

Table 2. Overview of selected features after fusion of dataset

Algorithm	**Selected features by applied feature selection algorithm**
RFE	HomeMaintainer, ECTS_1y, StudyDuration, AvgGrade, PointsInc
GA	ECTS_1y, PointInc, CodeSchool, Attendance, HomeMaintainer, StudyDuration, PointsInc, Gender, AvgGrade, Married, EdFather, Age
Boruta	StudyDuration, HomeMaintainer, HomeInc, EdFather, EdMother, Attendance, Married, PointsInc, AvgGrade, ParentEd, CodeSchool, Study
Fusion Dataset	HomeMaintainer, Attendance, ECTS_1y, StudyDuration, CodeSchool, AvgGrade, PointsInc, EdFather, HomeInc

Table 2 shows selected features by each algorithm regarding their importance. In the second step, accordantly with proposed method in [12], the five top and repeated features selected by any of the applied features selection algorithms were fused and applied in our final experiment.

2.4. Multiclasification modelling by random forest

The random forest algorithm (RF) is a machine learning algorithm widely used in classification and regression problems. It contains an ensemble of independent decision trees built on different data samples. Each tree is generated by randomly selected variable. The main advantages of random forests are high predictive accuracy and their applicability in high-dimensional problems with highly correlated variables. The variable importance measures obtained by random forests have also been suggested for the selection of relevant predictors in the analysis of student's dropout.

3. Results

During the research, the random forest algorithm was trained and tested on different dataset of selected features. The five classification models are compared by using the following well known evaluation measures for classification: overall accuracy, % of correctly classified instances, Accuracy, Precision, Sensitivity, and Specificity. According to Table 3, the random forest applied on the dataset of the fused features has the highest accuracy rate. The accuracy rate is addressed to the ratio of correctly estimated samples to the number of all samples. In our research, it is 85.15% for RF applied on fused dataset, which represent excellent results.

Table 3. Validation metrics: Accuracy of applied classifiers

Model		**RF_all fetures**	**RF+RFE**	**RF+GA**	**RF+Boruta**	**RF+fusion**
Accuracy		0.8023802	0.736271	0.8415842	0.8316832	0.8514851
Kappa		0.7006138	0.6369138	0.6383908	0.6505090	0.7227834
Accuracy	A	0,8138	0,8014	0,8138	0,8014	0,8229
	D	0,9325	0,9241	0,9325	0,9241	0,9552
	G	0,8934	0,8934	0,8934	0,8934	0,8934
Precision	A	0,7222	0,6894	0,6924	0,6656	0,701
	D	0,8359	0,7828	0,7901	0,7701	0,8139
	G	0,757	0,7996	0,7787	0,774	0,7927
Sensitivity	A	0,667	0,6782	0,674	0,6399	0,6816
	D	0,8379	0,8237	0,8249	0,8254	0,8469
	G	0,8478	0,8165	0,809	0,799	0,821
Specificity	A	0,8442	0,83	0,8301	0,817	0,8361
	D	0,9415	0,9221	0,9257	0,9201	0,9361
	G	0,8655	0,8822	0,8713	0,8676	0,8812

Table 4. Importance of features for the best model

Variable	**Importance**
StudyDuration	0.21185714286
ECTS_1y	0.09994285714
HomeMaintainer	0.01256571429
AvgGrade	0.00805714286
EdFather	0.00759428571
PointsInc	0.00628000000
Attendance	0.00465142857
CodeSchool	0.00108000000
HomeInc	0.00009714286

4. Discussion and Conclusions

While tracking the study performance at the Faculty of Maritime Studies University of Split, a large dropout is noticed. The analysis shows that a large number of students are studying for more than 3 years, which is above the expected time of graduation. To improve the performance of studying and faculty rating, in order to identify potential risks of dropouts and prevent students from leaving studies, a more fundamental approach should be used.

Results also show that students from other schools (except maritime, technical, and gymnasium) have a greater risk of dropping out. Furthermore, if their responses are linked with high school grades and the total number of points obtained in the final state exam, it is clear that a student with a mid-grade score of less than 3.52 and with a lower average number of points are dropping the most. The analysis indicates that the highest dropout rate is at the PTJM study. The students stated out that the main reason for dropping out is the transition to another study. The fact that they couldn't enroll in the desired study is also one of the reasons for a student dropping out. From the analysis, it is evident that these students generally come from other schools that are not

close to technical or maritime areas, or that they enroll in university only to meet their parent's expectations. Hence, the choice of study can be related with the student's background and family tradition.

Regarding duration of the study, the longest study duration is noticed on PN, PEIT, and BS. This could be connected to the fact that during the study period, these students usually start their internships on-board.

The obtained results of our research show that higher education institutions, especially the one where this research was conducted, should be aware at the earliest stage of the need to determine the profile of students with a potential risk of dropping out of school. Machine learning predictive modeling is of great benefit in systematically determining the risk of dropping out and diagnosing the cause depending on the level of risk. In this study, a random forest model combined with selected features obtained by fusion of different feature selection algorithms showed excellent performance in multi-class classification and student dropout prediction. Faculty Management together with Quality Assurance System should consider the possibility and need of building a dropout early warning and support systems for at risk students using the predictive models.

References

[1] Kehm, B. M.; Larsen, M. R.; Sommersel, H. B. Student Dropout from Universities in Europe: A Review of Empirical Literature. *Hungarian Educ. Res. J.*, **2020**. https://doi.org/10.1556/063.9.2019.1.18.

[2] OECD Indicators. *Education at a Glance 2020*; 2020. https://doi.org/10.1787/69096873-en.

[3] OECD. *Education at a Glance 2019*; 2019. https://doi.org/10.1787/f8d7880d-en.

[4] Ferreyra, M. M.; Avitabile, C.; Botero Álvarez, J.; Haimovich Paz, F.; Urzúa, S. *At a Crossroads: Higher Education in Latin America and the Caribbean*; 2017. https://doi.org/10.1596/978-1-4648-1014-5.

[5] Pierrakeas, C.; Koutsonikos, G.; Lipitakis, A. D.; Kotsiantis, S.; Xenos, M.; Gravvanis, G. A. The Variability of the Reasons for Student Dropout in Distance Learning and the Prediction of Dropout-Prone Students. In *Intelligent Systems Reference Library*; 2020. https://doi.org/10.1007/978-3-030-13743-4_6.

[6] Mesarić, J.; Šebalj, D. Decision Trees for Predicting the Academic Success of Students. *Croat. Oper. Res. Rev.*, **2016**, *7* (2), 367–388. https://doi.org/0.17535/crorr.2016.0025.

[7] Perchinunno, P.; Bilancia, M.; Vitale, D. A Statistical Analysis of Factors Affecting Higher Education Dropouts. *Soc. Indic. Res.*, **2019**. https://doi.org/10.1007/s11205-019-02249-y.

[8] Brownlee, J. Recursive Feature Elimination (RFE) for Feature Selection in Python. *Data Preparation*. May 25, 2020.

[9] Park, J.; Choi, E.; Kang, M.; Jung, Y. Dropout Genetic Algorithm Analysis for Deep Learning Generalization Error Minimization. *Int. J. Adv. Cult. Technol.*, **2017**, *5* (2), 74–81. https://doi.org/https://doi.org/10.17703/IJACT.2017.5.2.74.

[10] Queiroga, E. M.; Lopes, J. L.; Kappel, K.; Aguiar, M.; Araújo, R. M.; Munoz, R.; Villarroel, R.; Cechinel, C. A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. *Appl. Sci.*, **2020**. https://doi.org/10.3390/app10113998.

[11] Chen, R. C.; Dewi, C.; Huang, S. W.; Caraka, R. E. Selecting Critical Features for Data Classification Based on Machine Learning Methods. *J. Big Data*, **2020**. https://doi.org/10.1186/s40537-020-00327-4.

[12] Ahmed, A.; Malebary, S. Feature Selection and the Fusion-Based Method for Enhancing the Classification Accuracy of SVM for Breast Cancer Detection. *Int. J. Comput. Sci. Netw. Secur.*, **2019**, *19* (11), 55–60.