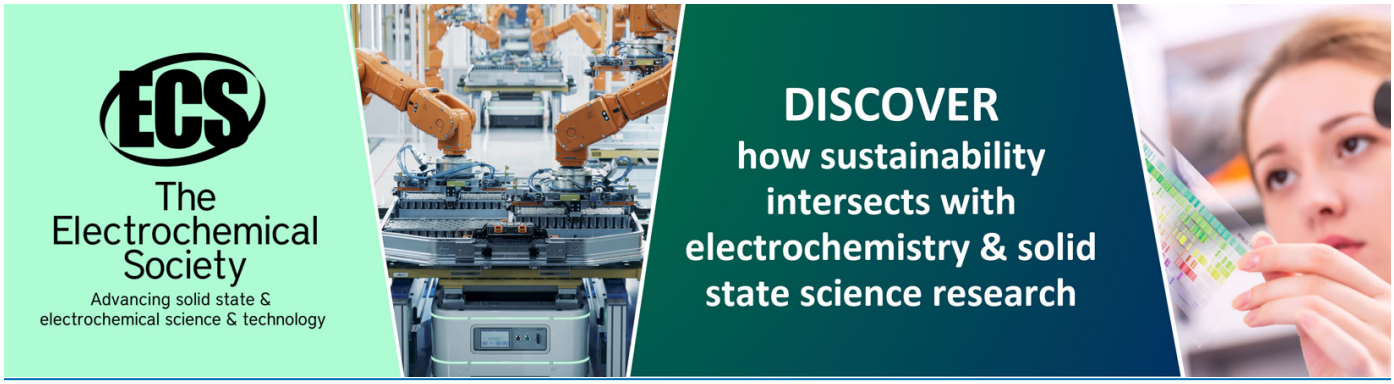
Attributes selection using machine learning for analysing students' dropping out of university: a case study



Attributes selection using machine learning for analysing students' dropping out of university: a case study

T I Pehlivanova and V I Nedeva

Trakia University of Stara Zagora, Faculty of Technics and Technologies of Yambol, Department of Electrical engineering, electronics and automatics, 38 Graf Ignatiev str. Yambol, Bulgaria

e-mail: tanya.pehlivanova@trakia-uni.bg

Abstract. Many students in Bulgarian universities drop out of the university before completing their studies. Identifying students at risk of dropping out allows timely taking measures for their retention. The paper presents the results of a study conducted among students of engineering programs at Trakia University - Stara Zagora. The collected data are subjected to processing, which aims to find the most important attributes that determine the risk of dropping out of university. The processing is done with Weka open source software. Different algorithms for selecting attributes with different search methods are applied. The most appropriate attribute selection algorithm was selected after applying the BayesNet classifier to the results obtained. The indicators TP rate, Precision and F-measure were compared. When applying InfoGainAttributeEval, the highest results are obtained for the accuracy of the classification. At the next stage, it is planned to expand the study among a larger number of students from different programs and create an effective forecasting model.

1. Introduction

Students' dropping out of university is an important problem for all countries around the world (including the most advanced).

In [1] a study is presented, according to which in the countries of the Organization for Economic Cooperation and Development OECD "by the beginning of the second year of study, an average of 12% of bachelor's degree students have left the tertiary education system. Only 39% of those who enter a bachelor's program graduate within the theoretical duration of the program; another 28% graduate during the following three years."

Identifying students at risk of failure or dropout is important for students, universities and countries. For students, failure leads to stress, limitations on their future career development, financial losses. For the institutions, in addition to the financial impacts, the dropout rate of students also affects the place in the rating rankings. And this leads to a decrease in students' interest in the respective university. In countries where the government funds all or part of higher education, society as a whole pays indirectly for student failure.

Universities can take various measures to reduce the probability of students dropping out before completing the program - financial incentives or sanctions, student support (counseling, career guidance, mentoring), organization of higher education (providing opportunities for individual learning, shortening of training, etc.) [2].

The aim of the paper is to find the most appropriate method for selecting attributes and to determine

the attributes that have the greatest impact on the dropout of students from the university. At the next stage, it is planned to expand the study among a larger number of students from different programs at Trakia University - Stara Zagora and to create an effective forecasting model.

2. Related work

Different types of data are used to apply the classification algorithms. There are studies in the literature in which the models are based only on personal and pre-university characteristics [3, 4, 5]. Other authors use data for the studied disciplines [6, 7]. In [8] the conclusions were made using socio-demographic indicators (age, date of birth, geographical location, marital status, parental education, parents' profession and annual income), educational factors (high school results, school location) , parental attitudes, attitude towards the university, etc.

In [9] a full analysis of the factors that lead to student's dropping out of university is made. 112 factors are indicated. These factors were classified into five dimensions: personal, academic, economic, social, and institutional. The conclusion shows that the most commonly studied was the personal dimension, which considers factors such as age, ethnicity and gender.

In [10] an extensive review of the literature related to the dropout of students in European universities is made and the reasons for dropout are summarized in 9 dimensions: (a) study conditions at university, (b) academic integration at university, (c) social integration at university, (d) personal efforts and motivations for studying, (e) information and admission requirements, (f) prior academic achievement in school, (g) personal characteristics of the student, (h) sociodemographic background of the student, and (i) external conditions.

There are also publications that discuss specific factors such as addiction to the internet, social networks and technologies [11].

Many attribute selection techniques can be applied to the same data. Different approaches are possible to determine the most appropriate one. Usually one or more classifiers are selected and their performance is compared. The comparison is performed by time [12], classification accuracy [12, 13], Precision, Recall, F-measure and Mean, absolute error [14].

The time indicator is not suitable for our study, as the processing time is short. The number of attributes is 32, and the number of instances is in line with the objectives of the current stage of researching the problem. The need for selection the most appropriate attributes arises from the fact that often the information collected by students is sensitive. It concerns data on marital status, income, education, grades, family stress, personal stress, etc., which respondents are not always willing to share. The indicators TP rate, Precision, F-measure were chosen as an indicator for the effectiveness of the classification in our study. The results show that by optimizing the number of characteristics, it is possible to increase the accuracy of predicting the risk of dropping out of university.

3. Methods

The data were collected through a survey among students from 2 engineering programs of TrU - Stara Zagora. The total number of surveyed students is 115. 91 of the students continue their studies and 24 have left the university for various reasons. The survey includes 32 questions. They are divided into three groups - personal characteristics, academic environment and social factors. The questions were compiled after a thorough review of the literature, which analyzes the reasons for dropping out of university. The survey was conducted in the period from 10.10.2019 to 01.20.2020.

Data pre-processing involved several procedures that can be represented as follows: Data accumulation; Data cleaning; Data transformation.

The final data set used to execute the project contains 115 instances, each described with 23 attributes with nominal variables.

Values of attributes are as follow:

@attribute 1.AGE {27-29,30-32,21-23,24-26,19-20}; Age

@attribute 3B.COURSE/YEAR OF EDUCATION {'Second Year','First Year','Third Year','Fourth Year'}; Course/Year of Education

@attribute '4.MAR_STAT ' {Single,Married,Other}; Marital status of students

@attribute 5.CHILD {'Have not children','Have children'}; Children of the student

@attribute 6.JOB {'yes, in the other specialty','Yes, in the specialty','I do not work'}; Does the student work in the specialty

@attribute '7.SATISF ' {'neither yes nor no ','rather no ',yes,'rather yes',no,'no opinion'}; Job satisfaction

@attribute '8.INCOME ' {medium,'low ',high,'very low ','very high'}; Family incomes

@attribute 9.PLACE {'small town','big city',Sofia,village}; Place of residence

@attribute'10.EDU_PARENT ' {'higher education','higher education- secondary education','secondary education','secondary education-primary education ','primary education ','higher education-primary education '}; Parental education

@attribute '12.PERS_STRESS ' {'yes - financial ','have not stress','problems with colleagues and learning problems','yes - illness','yes - problems with teacher','yes - learning problems ','yes - change in personal purpose','yes - other type','yes - change in personal purpose and problems with colleagues ','yes - financial yes - learning problems yes - problems with teacher','yes - illness yes - learning problems ','yes - change in personal purpose yes - problems with teacher','yes - learning problems yes - change in personal purpose','yes - learning problems yes - problems with teacher'}; Personal stress

@attribute '14.HIGH_SCH ' {'profiled high school (language mathematics or other)','vocational high school with different specialty','vocational high school in the specialty','non profiled high school'}; Profile of completed secondary education

@attribute 15.ASSES {excellent,'very good',good,middle}; Average success from high school

@attribute '20.SATISF_TRAIN ' {'rather yes',yes,'neither yes nor no','rather no',no}; Satisfaction with the level of education (overall)

@attribute '21.SATISF_CURRI ' {'rather yes',yes,'neither yes nor no','rather no',no}; Satisfaction with the subjects in the curriculum of the specialty

@attribute '22.SATISF_INFRA ' {'rather yes',yes,'rather no',no,'neither yes nor no'}; Satisfaction with the educational infrastructure (laboratories, dormitory, office, etc.) at the university

@attribute '23.SATISF_ADMIN ' {'rather yes',yes,'neither yes nor no','rather no',no}; Satisfaction with the administrative service of students

@attribute '24.SAT_REL_PROF ' {'rather yes',yes,'rather no',no,'neither yes nor no'}; Satisfaction with communication between teachers and students

@attribute '25.SAT_REL_STUD ' {yes,'neither yes nor no','rather yes',no,'rather no'}; Satisfaction with student relationships

@attribute '26.SATISF_DIFFIC ' {'rather yes','neither yes nor no',yes,'rather no',no}; Satisfaction with the difficulty and volume of the content of the curriculum in the subjects of the specialty

@attribute '27.SATISF_QUALIF ' {yes,'neither yes nor no','rather no','rather yes',no}; Satisfaction with opportunities for professional development

@attribute '28.SATISF_FUTU ' {'rather yes',yes,'rather no','neither yes nor no',no}; Graduation from university is a prerequisite for professional success in the future

@attribute 32.EDU_STATUS {active,dropout}; Student status

The collected data for the application of machine learning algorithms often contains attributes that are not relevant for making predictions. The attributes selection to remove and those to be involved in creating the model is a difficult task. Different algorithms are used to implement it.

Attributes selection is the process of reducing the number of input variables when developing a prediction model.

The reason for reducing the number of input attributes may be to reduce the computational cost of modeling and in some cases to improve the performance of the model.

Depending on whether the attributes are selected based on the target variable or not, the methods for attributes selection are supervised and unsupervised

Unsupervised feature selection techniques ignore the target variable, such as methods that remove redundant variables using correlation. Supervised feature selection techniques use the target variable, such as methods that remove irrelevant variables.

The article uses supervised methods. They are divided into:

- Filter Methods
- Wrapper Methods
- Embedded Methods

There are machine learning algorithms in which the feature selection algorithm is part of the model training. An example of an embedded technique is the decision tree algorithm.

Another way to reduce data processing time and resources is dimensionality reduction. In this case the input data is projected into a feature space with lower dimensions. This is different from attributes selection.

There are various techniques in the Weka platform for selecting machine learning datasets. For the purposes of this article, 6 of them are selected as the most suitable. A brief description of the selected algorithms is given in table 1 [15].

Table 1. Evaluators and their actions.

Evaluator	Action
InfoGainAttributeEval	Evaluates the worth of an attribute by measuring the information gain with respect to the class. InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute).
CfsSubsetEval;	Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of attributes that are highly correlated with the class while having low intercorrelation are preferred.
CfsSubsetEval; GreedyStepwise - GenerateRanking: True	Ranking is the order that attributes were added, starting with no attributes. The merit scores in the left column are the goodness of the subset after the adding the corresponding attribute in the right column to the subset. Attribute Subset Evaluator (supervised, Class (nominal): CFS Subset Evaluator
CorrelationAttributeEval;	Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.
GainRatioAttributeEval;	Evaluates the worth of an attribute by measuring the gain ratio with respect to the class. GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).
OneRAttributeEval	Evaluates the worth of an attribute by using the OneR classifier.

When applying attributes evaluation algorithms, there are several different search methods: BestFirst, GreedyStepwise and Renker. Their application also depends on the evaluation algorithm used. In each experiment, the algorithm for selecting the attributes (Evaluator) and the search method with its parameters are specified. The experiments were performed for Class 32.EDU_STATUS as the output attribute, and the other 22 attributes were ranked. The algorithm also determines the coefficients according to which the ranking is made.

To determine the optimal subset of attributes for each algorithm, different approaches are possible: The most commonly used approach is to set a limit value of the coefficient and all attributes that have a

lower value to be cut (removed from the database). In our study, we used the following approach: An appropriate classifier was selected and a classification was made with different subsets of attributes. They are defined as follows: First, the lowest ranked attribute is removed. After that the two lowest ranked attributes are removed and so on.. This subset is selected in which the accuracy of the classification is the highest.

In order to determine the optimal selection algorithm, classification is made with the selected optimal subsets of attributes for each algorithm. The indicators from the classification are compared. This algorithm is selected, in which the indicators TP rate; Precision; F-Measure have the highest values.

4. Results

The first task that the research solves is to make the selection of the attributes with all the algorithms described in item 3. As a result, the attributes are ranked by significance (table 2). In addition to the result of the execution of the algorithms, the table also contains the options for their application.

For example, in Selection 1, the algorithm that is applied is InfoGainAttributeEval. The search method in this case is for ranking - Ranker -T -1.7976931348623157E308 -N -1; Information Gain Ranking Filter. The result obtained from all 22 analyzed attributes is with the following order - 12, 1, 3B, 24, 10, 7, 4, 25, 20, 14, 28, 6, 27, 21, 8, 9, 26, 5, 23, 22, 15, 2.

Table 2. Ranked attributes in the different selection methodologies.

Selection N, Attribute Evaluator, Search method and Options	Ranked attributes
Selection 1: InfoGainAttributeEval Ranker -T -1.7976931348623157E308 -N -1; Information Gain Ranking Filter	**12,1,3B,24,10,7,4,25,**20,14,28,6,27, 21,8,9,26,5,23,22,15,2
Selection 2: CfsSubsetEval; BestFirst -D 1 -N 5;	**1,3B,10,12,24**
Selection 3: CfsSubsetEval; GreedyStepwise -R -T -1.7976931348623157E308 -N -1 -num-slots 1; GenerateRanking: True Including locally predictive attributes	**2,1,3B,10,24,4,25,14,6,20,9,8,5,28,7, 27,2,22,**26,21,15,23
Selection 4: CorrelationAttributeEval; Ranker -T -1.7976931348623157E308 -N -1; Correlation Ranking Filter	**1,12,3B,5,24,14,4,6,9,25,8,10,7,21, 28,22**,20,27,15,23,26,2
Selection 5: GainRatioAttributeEval; Ranker -T -1.7976931348623157E308 -N -1; Gain Ratio feature evaluator	**12,1,3B,24,10,4,25,20,5,6,14,28,7,8, 27,9,21,26,**23,22,15,2
Selection 6: OneRAttributeEval OneRAttributeEval -S 1 -F 10 -B 6 Ranker -T -1.7976931348623157E308 -N -1 Using 10 fold cross validation for evaluating attributes. Minimum bucket size for OneR: 6	**1,12,26,15,6,5,2,7,10,20,22,24,23, 21,8,14,27,4,25,**9,3B,28

The methodology described in item 3 is applied to the results obtained for each algorithm and the optimal attributes are determined.

The comparison of the different subsets of attributes was made using the BayesNet classifier. This classifier is chosen because it gives the best results according to the nature of the data collected. The choice of classification algorithm was made at a previous stage of the study. The output class is Output 32.Edu_Status. The approach that is applied is Full training set, because with it the results for accuracy and all other indicators are the highest.

The results of the application of the methodology to determine the optimal subset of attributes for the InfoGainAttributeEval algorithm, Selection 1, are shown in table 3. The best values of the

classification indicators are obtained for a subset consisting of 8 attributes, namely, 12,1,3B, 24,10, 7,4,25. In the same way the data from the other selections, from 2 to 6, are analyzed. In all selections a different optimal number of attributes is obtained. The selected attributes are shown in table 2 in bold.

Table 3. Selecting the optimal subset of attributes in Selection 1 InfoGainAttributeEval.

Number of attributes	Included Attributes	TP Rate	Precision	F-Measure
22	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9,26,5,23,22,15,2	0.904	0.901	0.902
21	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9,26,5,23,22,15	0.904	0.901	0.902
20	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9,26,5,23,22	0.896	0.893	0.894
19	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9,26,5,23	0.896	0.893	0.894
18	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9,26,5	0.896	0.893	0.894
17	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9,26	0.896	0.892	0.892
16	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8,9	0.887	0.883	0.884
15	12,1,3B,24,10,7,4,25,20,14,28,6,27,21,8	0.896	0.893	0.894
14	12,1,3B,24,10,7,4,25,20,14,28,6,27,21	0.878	0.875	0.876
13	12,1,3B,24,10,7,4,25,20,14,28,6,27	0.896	0.896	0.896
12	12,1,3B,24,10,7,4,25,20,14,28,6	0.887	0.885	0.886
11	12,1,3B,24,10,7,4,25,20,14,28	0.861	0.857	0.859
10	12,1,3B,24,10,7,4,25,20,14	0.887	0.885	0.886
9	12,1,3B,24,10,7,4,25,20	0.896	0.892	0.890
8	**12,1,3B,24,10,7,4,25**	0.913	0.910	0.910
7	12,1,3B,24,10,7,4	0.887	0.883	0.884
6	12,1,3B,24,10,7	0.878	0.872	0.872
5	12,1,3B,24,10	0.878	0.873	0.869
4	12,1,3B,24	0.878	0.873	0.869

The second task to be solved is to compare the results obtained from the application of different algorithms. Solving this problem will determine the optimal number and composition of attributes.

The results obtained from the application of the BayesNet classifier for the optimal subset of attributes for each of the selections were used. The comparison is again made on the indicators TP rate; Precision; F-Measure. The results are shown in table 4.

Table 4. Comparison of the classification indicators for different selections.

Attribute Selections	Number of Attributes	TP Rate	Precision	F-Measure
Classification with all attributes	***22***	***0.904***	***0.901***	***0.902***
Selection 1: InfoGainAttributeEval	8	0.913	0.910	0.910
Selection 2: CfsSubsetEval BestFirst	5	0.878	0.873	0.869
Selection 3: CfsSubsetEval	18	0.887	0.883	0.884
Selection 4: CorrelationAttributeEval	16	0.904	0.903	0.904
Selection 5: GainRatioAttributeEval	18	0.896	0.893	0.894
Selection 6: OneRAttributeEval	19	0.904	0.903	0.898

The analyses made give us reason to draw the following conclusions:

- The highest classification accuracy is obtained with selection 1 InfoGainAttributeEval. The classification indicators are TP Rate - 0.913; Precision - 0.910; F-Measure - 0.910.
- The optimal subset includes 8 attributes. With these attributes the following analyzes with large volumes of data about the faculty and the university will continue.
- The list of attributes to which special attention should be paid is given below. They are ranked according to InfoGainAttributeEval.

12.PERS_STRESS - This attribute contains the answers to the question about the personal stress that each student experiences as a reason for dropping out of university.

1.AGE - The age of the student is also important as a factor for dropping out.

3B.COURSE - The student's year of study greatly influences the interruption of his / her studies. The majority of students drop out of the first or second year of study.

24.SAT_REL_PROF - The relationship between students and lecturers can be a significant reason for dropping out.

10.EDU_PARENT - It turns out that parents' education influences students dropping out of university.

7.JOB_SATISF - When students work, job satisfaction has a significant impact on dropping out of university.

4.MAR_STAT - Marital status also matters.

25.SAT_REL_STUD - The answers to this attribute show satisfaction with the relationship with other students at the university. This also influences the decision to interrupt training.

5. Conclusion

The article presents the results of application of various methods for selection of attributes. Data from a study of the reasons for dropping out of university were used. The processing is done with the Weka software. To determine the optimal method for attribute selection, the BayesNet classifier was applied to the resulting datasets. The accuracy of the classification is compared. TP rate, Precision, F-measure were used as indicators. For the data used, the best selection of attributes is obtained with InfoGainAttributeEval. As a result of the study, the most important factors influencing the final decision of students to leave the university were identified.

At the next stage we plan to cover all students studying at the Faculty of Technics and Technologies - Yambol. The study will include only those identified as the most important factors. A recommendation will be made to the university management to keep up-to-date information for each student, to analyze these factors annually, to make a forecast for students who are at potential risk of dropping out and to take measures for their retention.

References

[1] OECD 2019, *Education at a Glance 2019: OECD Indicators* (Paris: OECD Publishing)
[2] Vossensteyn H, Kottmann A, Jongbloed B, Kaiser F, Cremonini L, Stensaker B, Hovdhaugen E and Wollscheid S 2015 *Dropout and Completion in Higher Education in Europe: Main Report* (Luxembourg: Publications Office of the European Union)
[3] Kabakchieva D 2013 Predicting Student Performance by Using Data Mining Methods for Classification *Cybernetics and information technologies* **13** pp 61-72
[4] Martinho V, Nunes C and Roberto C 2013 An Intelligent System for Prediction of School Dropout Risk Group in Higher Education Classroom based on Artificial Neural Networks *Proc. Int. Conf. on Tools with Artificial Intelligence (Herndon, VA, USA)* pp 159-166
[5] Dekker G, Pechenizkiy M and Vleeshouwers J 2009 Predicting Students Drop Out: A Case Study *Proc. Int. Conf. on Educational Data Mining (EDM'09) (Cordoba, Spain)* pp 41-50
[6] Askinadze A and Conrad S 2019 Predicting Student Dropout in Higher Education Based on Previous Exam Results *Proc. Int. Conf. on Educational Data Mining (EDM 2019) (Montréal, Canada)* pp 500-503
[7] Askinadze A and Conrad S 2017 Application of the dynamic time warping distance for the student drop-out prediction on time series data *Proc. Int. Conf. on Educational Data Mining (EDM 2017). (Wuhan, Hubei, China)* pp 342-343
[8] Rai S and Jain A 2013 Students' Dropout Risk Assessment in Undergraduate Courses of ICT at Residential University – A Case Study *International Journal of Computer Applications* **84** pp 31-36
[9] Alban M and Mauricio D 2019 Predicting University Dropout through Data Mining: A Systematic Literature *Indian Journal of Science and Technology* **12**
[10] Kehm B, Larsen M and Sommersel H 2019 Student dropout from universities in Europe: A review of empirical literature *Hungarian Educational Research Journal* **9** pp 147–164
[11] Alban M and Mauricio D 2018 Prediction of university dropout through technological factors: a case study in Ecuador *Espacios* **39** pp 8-15
[12] Kumara C, Sooraj M and Ramakrishnanc S 2017 A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets, *Proc. Int. Conf. on Advances in Computing & Communications ICACC-2017 (Cochin, India)* pp 209-217
[13] Phyu T Z and Oo N N 2016 Performance Comparison of Feature Selection Methods, *Proc. MATEC Web of Conferences 42*
[14] Gnanambal S, Thangaraj M, Meenatchi V and Gayathri V 2018 Classification Algorithms with Attribute Selection: an evaluation study using WEKA *Int. J. Advanced Networking and Applications* **9** pp 3640-44
[15] https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html