# ANALYSIS OF STUDENT DROPOUT AT WESTERN SERBIA ACADEMY OF APPLIED STUDIES: A CLUSTERING BASED APPROACH

**Milovan Milivojevic[1]; Srdjan Obradovic[2]; Ana Kaplarevic Malisic[3]; Slobodan Petrovic[4]**
[1] Western Serbia Academy of Applied Studies, Užice, Serbia, mmilivojevic031@gmail.com
[2] Western Serbia Academy of Applied Studies, Užice, Serbia, srdjan.obradovic665@gmail.com
[3] Faculty of Science, University of Kragujevac, 34000 Kragujevac, Serbia, ana@kg.ac.rs
[4] Western Serbia Academy of Applied Studies, Užice, Serbia, slobodan.petrovic@vpts.edu.rs

**Abstract:** Early prediction of student dropout in higher education and identification of factors which contribute the most to student attrition are necessary prerequisites in taking steps towards creating effective attrition prevention measures. Retaining students at higher education level is especially important in Western Serbia, where the number of students enrolled in academies of applied studies has a downward trend. Based on the data from the Western Serbia Academy of Applied Studies, we employed several statistical and data mining methods to perform various analysis of student dropout at this higher education institution. After preliminary exploration and data preparation, we examined relationships between variables by correlation analysis and $\chi 2$ tests of independence. To simultaneously analyze quantitative and qualitative variables, and explore similarities and groupings of students considering the dropout variable, we used dimensionality reductions methods — FAMD, t-SNE, and UMAP. The previous analyses were the basis for engineering and selection of features that served as input to various clustering methods. After cluster analysis for fitted models, we adapted them into predictive models for student dropout. These predictive models were then evaluated on the test dataset using relevant classification metrics, and on that ground, we chose the k-medoids model as best for predicting student dropout.

***Keywords:*** *student dropout, higher education, machine learning, clustering*

## 1 INTRODUCTION

Student dropout in higher education is a notable concern in many countries, with a growing amount of scientific literature addressing this problem [1–3]. Reduction of student attrition rates is necessary for achieving sufficient levels of educational attainment in a population. Scott et al. suggested in [4], that universities with high attrition or dropout rates may face the significant loss of tuition fees and potential alumni contributions. There are numerous studies showing that a sizable chunk of student dropout occurs in the first year of higher education. According to the research in [5], more than 50% of the student attrition can be attributed to the freshmen. Therefore, it is essential to identify students who are in risk of dropping out as early as possible. This is crucial for the economic progress of any country, because advanced economies all gear towards the knowledge-based economy [6]. The topic of preventing student dropout has also gained increased attention due to COVID-19 pandemic forcing many countries to move much of their education activities online, which is adding more complexity to an already multifaceted phenomenon [7].
In Western Serbia, retaining higher education students is especially important in the applied studies domain because the number of students enrolled in academies of applied studies has a downward trend over the past 5-10 years, which is shown in Fig. 1 [8,9].

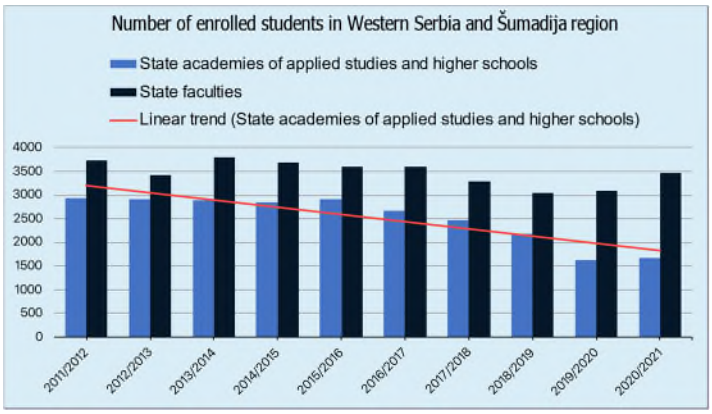



**Figure 1.** Number of students enrolled in higher education studies in Western Serbia and Šumadija region

Identification of students which are likely to drop out by using predictive models can be a first step in improving retention policies such as studying help, or mentorship aided learning. Predicting dropout may also be helpful for efficient distribution of pedagogical, psychological, and administrative resources.

Utilization of machine learning (ML) methods for predicting student dropout has also received great attention in recent times [10–12].

Student dropout prediction is closely related to "churn" (customer attrition) prediction, which is a task common in customer relationship management, and is solved by classification models [13,14]. Customer churn is costly and preventable. Businesses try to decrease customer churn by using predictive models which recognize customers at risk of churning and then take steps in improving business-customer relationships with them.

Student dropout prediction can be accomplished by using unsupervised and supervised classification. One form of unsupervised classification is clustering. In this paper we focus on using clustering models for student dropout prediction. Various clustering methods are used to model student dropout. Palani et al. [15] analyze student dropout as a major challenge in Virtual Learning Environments (VLE). Based on behavioral patterns in VLEs (student's engagement and academic performance), the authors developed a clustering model able to identify low student engagement early on in course progression. Experimental results showed that k-Prototypes clustering algorithm was more successful in finding segments of low VLE engagement students in comparison to Fuzzy C-means clustering, Gaussian Mixture Models, and Hierarchical models. The authors of [16] used different clustering validation metrics like the Dunn Index, Silhouette score and Davies-Bouldin score, to evaluate seven clustering models trained on a small dataset consisting of numerical grades for 336 students. The authors decided that k-means and PAM were the best performers among partitioning algorithms, and DIANA was the best performer among hierarchical algorithms. In [17], the authors showed that the clustering approach for student dropout can be of use for exploration of different student segments and their traits. To investigate student dropout from Brazilian universities, where close to 50% students drop out, Macedo et al. [18] employed the Fuzzy C-Means algorithm and applied the transactional distance theory to select features which were utilized in the clustering process. One major challenge in predicting undergraduate student dropout is the multidimensionality of data. The author in [19] focuses on dimensionality reduction of a student survey results dataset consisting of 51 five-point Likert scale numerical attributes which describe student behavioral features for 150 students. The goal was to identify major factors affecting student dropout by means of Principal Component Analysis (PCA). The author marked first 25 components out of 51 as important, and notes that a reduced dimensionality dataset is better for developing predictive models in comparison to the raw survey results dataset.

In Serbia, there is also a growing corpus of studies analyzing student dropout in higher education, both at university level [20], and applied studies level [21]. These studies consider the dropout phenomenon from various viewpoints, but to the best of our knowledge, we have not found studies employing statistical or machine learning models for predicting student dropout in Serbian higher education.

The goal of this paper is to make a contribution that would alleviate the consequences of the negative trend shown in Figure 1. Specifically, the goal is to choose a strategy, based on clustering, dimensionality reduction, and the premises of Learning Analytics, which would enable early detection of students at high risk of dropping out.

Our hypothesis is that it is possible to create a clustering-based model of student dropout which generalizes well, i.e., has sufficient predictive performance for new higher education students.

We explored the following clustering models on a training dataset: k-medoids, k-prototypes, hierarchical agglomerative, and two-step clustering [22]. To determine the optimal number of clusters for each clustering model we used various clustering validation metrics. The predictive models based on these clustering models were evaluated on a test dataset by appropriate classification metrics. Both, the training and testing dataset were prepared from raw student enrolment data, and their first-year performance data. The process also involved feature selection aided by dimensionality reduction. All the data was provided by the Western Serbia Academy of Applied Studies, Uzice department.

## 2 RESEARCH METODOLOGY

To extract meaningful insights from data, we defined a methodology based on the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [23] — a tried-and-true approach for data mining efforts.

The steps of our methodology are shown in Fig. 2.

The utilized methodology implies the preliminary exploration of data in the first phase (phase A). This involves getting to know the structure of the raw data — including number of rows, columns, character encoding method used, encoding of missing values, format of dates and time, performing various visualizations and summaries of the data, etc. After preliminary exploration, Data preparation steps follow which primarily consist of Data cleaning and Data normalization and encoding. In the data cleaning phase, the following is dealt with: Firstly, the problem of missing data is addressed. In this step (Fig 2., step A.2), the mechanisms of missing data [24,25] — missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) — are investigated. This involves using visualizations, t-tests for groups defined by missing versus non-missing values, and Little's missing completely at random test [26]. Based on these tests and domain knowledge, it is decided if, and how to perform imputation for each variable/column. In the next step of data cleaning (Fig 2., step A3) outliers for numeric variables are identified and then fixed, either by removal of whole rows, or imputation by method deemed appropriate. Categorical data outliers are defined as rare, erroneous, or nonsensical occurrences of a categorical variable. They are replaced with the appropriate category or, in the case of rare occurrence categories lumped together into a new category. In the normalization and encoding step (Fig 2., step A4) the following is carried out: (1) Normalization of numeric predictors. (2) Encoding of categorical predictors. (3) Encoding of the dependent variable (dropout).

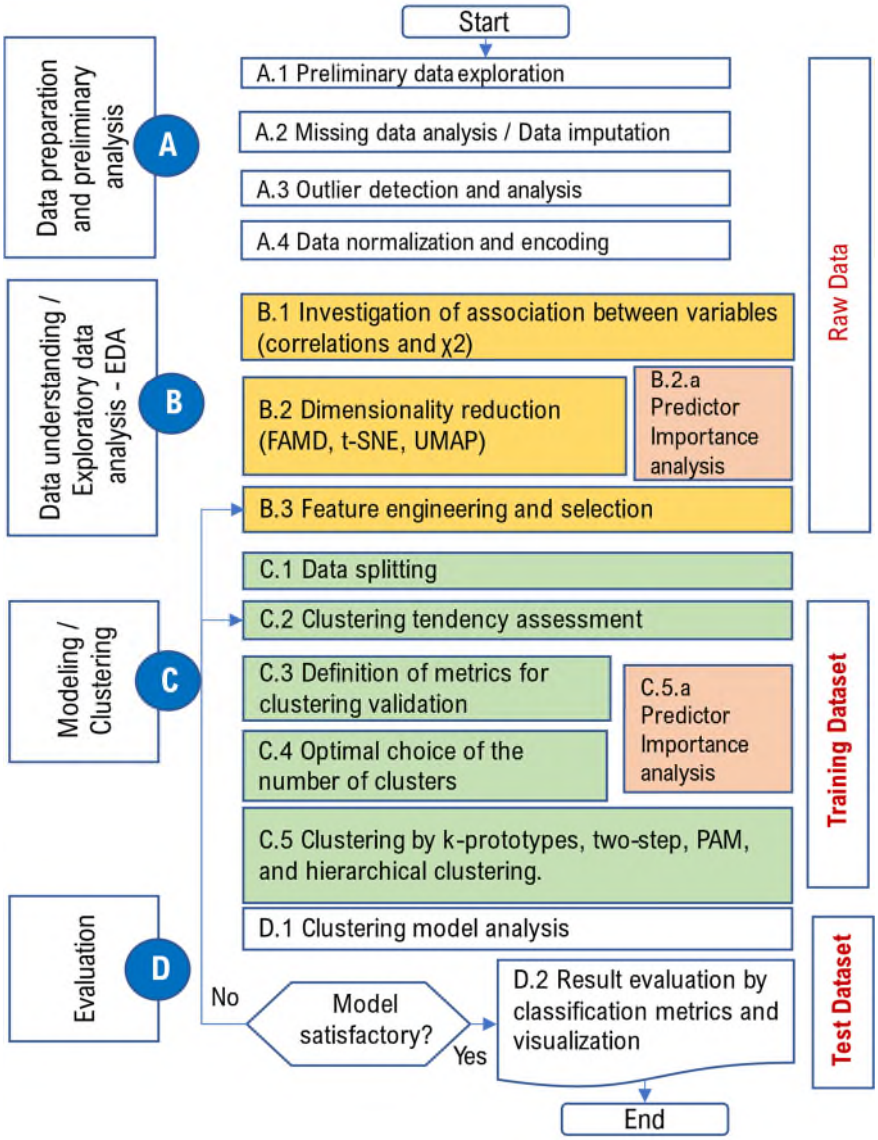



Fig 2. Employed methodology phases and steps

Within exploratory data analysis (EDA) (Fig. 2, phase B), the proposed methodology involves the following steps: Correlation analysis, Analysis of dependence for categorical variables, Dimensionality reduction, and Feature engineering/selection. Association between variables is investigated using Pearson correlation coefficients and scatter plots for numeric variables, and bar plots and $\chi 2$ test of independence for categorical variables (Fig. 2, step B1).

UMAP [27], t-SNE [28], and Factor analysis of mixed data method (FAMD) [29] dimensionality reduction is conducted to explore relationships between individuals by 3D visualization (Fig. 2, step B2). Additionally, FAMD is used to analyze relationships between both categorical and numeric variables, and provide predictor importance (Fig. 2, step B2.a). FAMD is suitable for dimensionality reduction and exploration of mixed data type datasets. It can analyze relationships between both categorical and numeric variables. The dimensionality reduction it offers also enables exploring similarities between individuals by visualization in 2D/3D space. This analysis has the purpose of aiding decision making in the next phase of feature engineering.

Feature engineering and selection (Fig. 2, step B3) is conducted based on steps B1, B2, and domain knowledge. Spurious columns/variables (i.e., variables irrelevant to the modeling of the student dropout phenomenon, or highly correlated/redundant with other variables) are identified and removed. Additional features are introduced and derived from existing features to describe students better in context of the dropout phenomenon.

In phase C the data are first split into training and testing datasets according to the defined ratio and time periods (Fig. 2, step C1). Clustering tendency assessment (Fig. 2, step C2) for the dataset is performed by calculating the Hopkins statistic [30] and inspecting visualizations produced in phase B for meaningful groupings. The Hopkins test statistic measures the probability that a given data set is generated by a uniform distribution.

Metrics for internal clustering validation are defined for each clustering method to determine the optimal number of clusters, and comparison between clusterings (Fig. 2, step C3). Metrics: Silhouette index, Within-cluster sum of squares, Dunn index, Calinski and Harabasz index, Average within/between distance ratio, and Entropy of the distribution of cluster membership [31], were used for clustering validation of Two-step clustering, K-medoids (partitioning around

medoids), and Hierarchical agglomerative clustering. On the other hand, metrics: Silhouette index, McClain index, Gamma Index, Gplus Index [32] were used for clustering validation of K-prototypes clustering.
Optimal choice of the number of clusters, K, is determined based on defined metrics and clustering models in C3 (Fig. 2, step C4).
Clustering on the training dataset is performed for each clustering models: Two-step clustering, K-prototypes, K-medoids, Hierarchical agglomerative clustering (Fig. 2, step C5). Two-step and K-prototypes use mixed variable type datasets as input. The remaining clustering methods instead use dataset dissimilarity matrices based on the defined distance measure. Two-step clustering method also provides predictor importance analysis [22] which can be useful for further insight into the student dropout phenomenon.
Finally in phase D, for all clustering models, the cluster with most dropped out students is chosen as a representative "dropout cluster" for that model. Then, cluster labels with that cluster number serve as "dropped out" labels, and all other cluster labels serve as "not dropped out" labels. With this scheme, confusion matrices and metrics — precision, recall, specificity, accuracy, balanced accuracy — are produced for both the test and training dataset, for all cluster models. If the classification metrics for the test are not satisfactory, corrections are made by making corrections in step C4, and/or step B3.

# 3 IMPLEMENTATION

## 3.1 Data preparation

### *3.1.1 Dataset*

The dataset data was provided by the Western Serbia Academy of Applied Studies (WSA), Uzice department, which is one of the academies in Western Serbia region. There are 5216 students in this dataset for 285 different courses classified in nine study programs, for the period from 2007 to 2020. The data was extracted from the Student Service Information System (SSIS) based on Microsoft Access RDBMS, and exported into several CSV files by appropriate SQL queries. The dataset contains anonymized data describing four different types of information related to student's (1) demographic data, (2) assessment scores from high school, (3) grades in entrance exams, and the (4) student's exam taking and semester enrollment history. Due to the fact that students transferring from other higher education institutions exhibit a different pattern of exam taking history and semester enrollment, they are not included in the analysis. Also, due to the time frame necessary for the test data, students from the Health care study program were also excluded from analysis because this program was first introduced in 2018. For these reasons the dataset was reduced to 3559 records, covering the period from 2007-2017. The supplied CSV files were transformed and prepared using the R programming language and tools from the *tidyverse* set of of R packages [33]. This resulted in an aggregated table where each row represented one student, and all available relevant information for that student. This table, as previously stated contained four groups of variables (columns in the table): student's demographic data {anonymized student ID, year of enrollment, year of birth, gender, region of birth in the Republic of Serbia, Age of student at enrollment date, etc.}, assessment scores from high school {Secondary education category, Secondary education duration, Place/Region of high school, Achieved points from high school education}, grades in entrance exams{Points on the math/biology/economy test, Points on the general culture test}. The student's exam taking and semester enrollment history is described by two groups of variables. The first group consists of summary values for the indicators of student's success: {mean grade in all exams, duration of studies, total number of passed exams, total number of tries for all exams, etc}. The first group also included categorical variables which served as indicators of important events during the studies {First30, First48, Third37, etc.}. For example, the variable which indicates if the student achieved 37 ECTS in the school year — *First37*. The second group included variables that describe the dynamics of student exam taking performance, and semester enrollment history. Based on the fact that the maximum number of years of study is seven, each of possible seven years of study is viewed as a separate entity. For each such entity (year of study) there are two types of variables: raw and derived. For example, for the entity *third year of stydy*, the raw variables are (1): ordinal number of enrolled semester, student's funding status for the third year (budget financed, self-financed, frozen), the number of passed exams in the third year, the number of achieved ECTS points in the third year, etc. The derived variables are (2): mean grade for the third year, incremental achievement of ECTS points, etc. This scheme is applied for each year of each student's studies.
Next, the conditions under which the student is marked as "dropped out" are defined. The students is labeled as "dropped out" if he hasn't graduated after four years of study. The dataset isn't imbalanced (which is often the case with the churning and dropout problems), and so methods for dealing with this were not needed.
The preliminary analysis extracted 85 potential predictors, and one dependent variable. Due to the large number of these predictors, they aren't described in detail here.

### *3.1.2 Missing value analysis, Outlier analysis, Data normalization and encoding*

Additionally, in the data preparation step the following is examined: distributions of variables, missing values (step A2, Fig. 2) and outliers. (step A3, Fig. 2).
As described in section 2.2, missing data mechanism (MCAR, MAR, and MNAR) were investigated by using visualizations, t-tests, $\chi^2$ tests, and Little's MCAR test. Visualizations were firstly used to investigate associations between

missingness and observations, and spot relationships of missingness between variables. Little's MCAR test was used to assess if data is MCAR. Additionally, t-tests and $\chi^2$ tests were carried out between each variable with missing values and other variables in the data set to see if the percentage of missing values for the investigated variable differs for different values of another variable. Missingness dummy variables were created to indicate whether an examined variable is missing (with value of 1 if an observation for the examined variable is missing, and a value of 0 if it is not).
For numerical variables, independent samples t-test was performed to see if there a difference in means for different levels of the dummy variable. If any of these two tests was significant ($p<0.01$) we concluded that there is a MAR mechanism for the examined variable. Based on the identified missingness mechanism for each variable with missing values, we decided whether to impute the missing data for that variable, or delete the whole observation, i.e. row of data. If the identified mechanism was MAR we imputed the values, because deleting observations introduces bias in clustering models we built on this data [24,25].
The results of the analysis and the imputation itself are not shown due to space limitations.
The missing values for the numerical variables were imputed in the following way: for *mean grade on exams* – 5; for *number of passed exams* – 0; for *points on the math/biology/economy test* – 10; for *points on the general culture test* – 10; For categorical variables such as *Secondary education category*, *Region of birth*, etc. we assigned the value of *other* for missing values.
No outliers were observed during the analysis, which indicates that the SSIS has adequate validation rules for the data.
The numeric data was normalized using min-max normalization method ([0-1] range) which is recommended for the clustering methods we used. In the next step, to enable reduction of dimensionality by UMAP and t-SNE methods data encoding of categorical variables was performed. The categorical variables were encoded using C-1 dummy encoding scheme. For all levels C of the categorical variable, C-1 dummy variables are introduced (Fig 2., step A4).

### 3.2 Data understanding

In phase B, exploratory data analysis (EDA) was performed for the selected aggregated dataset.

#### *3.2.1 Correlation and $\chi^2$*

Initially, correlation analysis was performed, including tests for significance of the Pearson correlation coefficients, *r*. Due to the large number of variables, the correlation matrix is shown only for selected numerical variables (Fig.3).

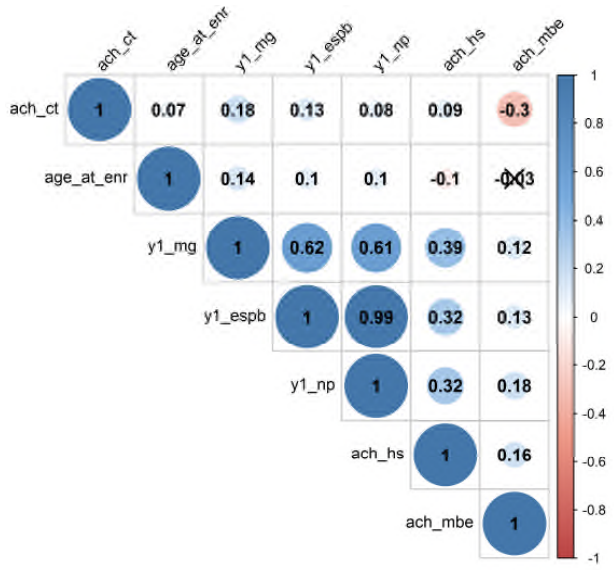



Fig 3. Correlation matrix for the selected variables

Based on the obtained small values of Pearson's *r* coefficients, it can be concluded that the predictor variables are not strongly linearly correlated, and thus the problems of pronounced multicollinearity are not present. The exceptions are variables *y1_ects* and *y1_np* which are strongly correlated (0.99), because they reflect the same phenomenon. Both of these variable are retained in further analysis for cases where achieved ECTS points differ significantly compared to most courses.
To gain a deeper insight into the bivariate associations between the selected categorical variables a series of $\chi 2$ independence tests were performed. The results of these tests are shown in Table 1, and the clustered bar plots for the selected cross tabulated variables are shown in Fig. 4.

From Fig. 4a it can be seen that the *gender* variable is significantly associated with student dropout, and that women drop out of studies significantly less than men.

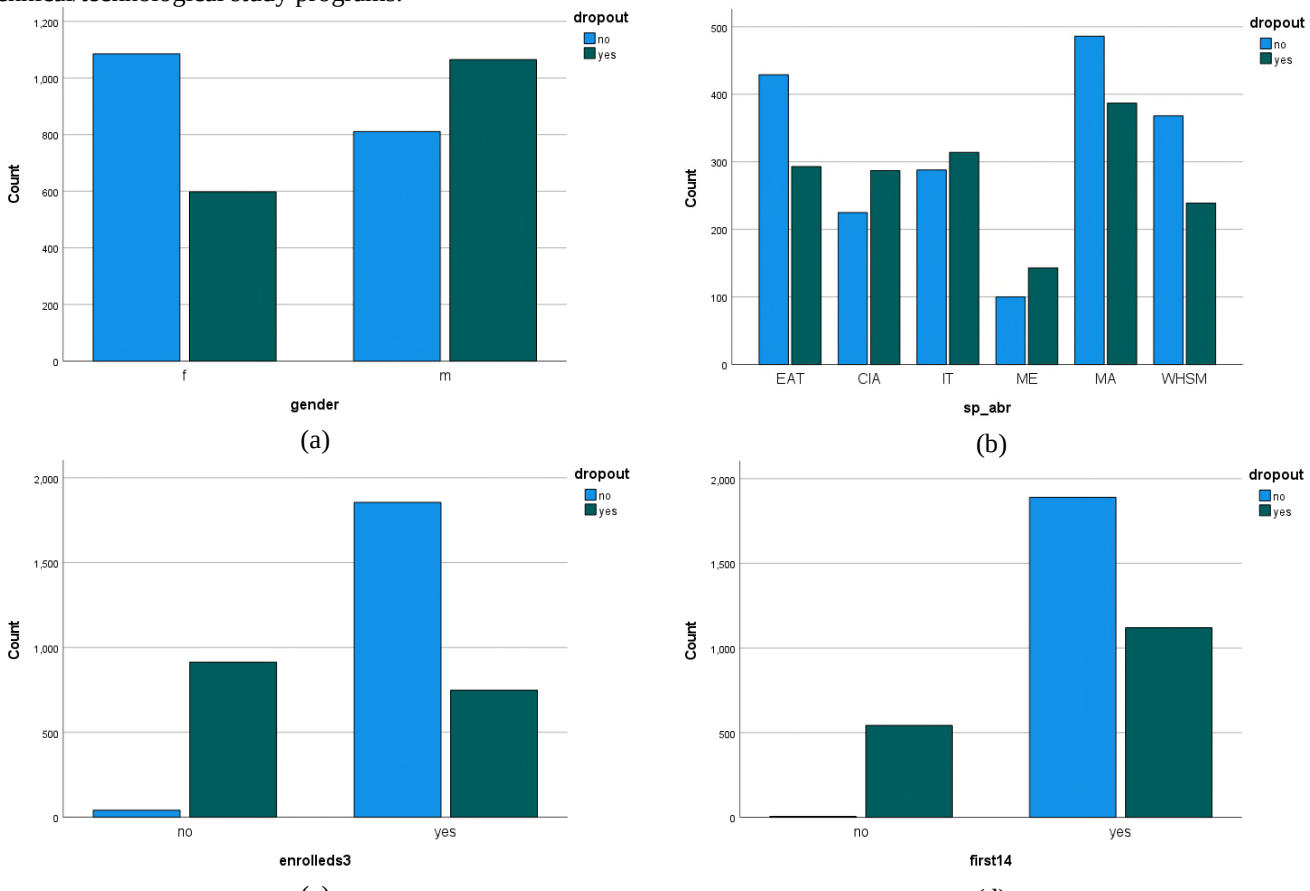
Due to the substantial number of pairs of variables that could potentially be analyzed, and the limits in space, a descriptive comment is given for some of these pairs. The analysis shows that the categorical variable *duration_hs*, is significantly associated with student dropout and that students who enrolled from three-year high schools, drop out much more compared to students who completed one of the four-year high schools. Also, after analyzing the impact of the high school category (variable: *cat_hs*), it can be seen that students who enrolled from technical high schools have a higher dropout rate.

Table 1. Tests of independence between categorical variables and the dropout variable

| n=3559 | χ2 | df | p | phi | Cramer's V | Conclusion |
|---|---|---|---|---|---|---|
| *Gender* | 159.89 | 1 | 1.19E-36 | 0.211 | | * a significant association |
| Category name of the study program (*abr_sp*) | 65.53 | 5 | 8.72E-13 | | 0.136 | a significant association |
| Enrolled into third semester (*enrolled3*) | 1255.24 | 1 | 6.02E-275 | -0.595 | | * a significant association |
| Achieved 30 ECTS in first 2 semesters (*first30*) | 1003.84 | 1 | 2.63E-220 | -.532 | | * a significant association |
| Achieved 37 ECTS in first 2 semesters (*first37*) | 1171.47 | 1 | 9.63E-257 | -0.574 | | * a significant association |
| Achieved 48 ECTS in first 2 semesters (*first48*) | 1205.36 | 1 | 4.16E-264 | -0.582 | | * a significant association |
| Achieved 14 ECTS in first 2 semesters (*first14*) | 707.55 | 1 | 6.81E-156 | -0.443 | | * a significant association |
| Region of birth (*reg_birth*) | 18.90 | 6 | 0.00432 | | .073 | a significant association |
| Secondary education category (*cat_hs*) | 46.62 | 5 | 6.77E-9 | | .114 | a significant association |
| Place/Region of high school (*place_hs*) | 45.39 | 12 | 0.000009 | | .113 | a significant association |
| Secondary education duration (*duration_hs*) | 26.96 | 1 | 2.07E-7 | 0.088 | | * a significant association |

A χ2 test for independence (with *Yates' Continuity Correction for 2x2 cross-tabs) indicated a significant association between selected categorical variables and *dropout*.

In addition, from Fig. 4b, it can be seen that students from study programs belonging to the technical/technological field (MA-Mechanical Engineering, IT-Information Technology, and GRiA-Civil Engineering and Architecture) dropout more than students who enrolled in a study program from the field of social sciences. We believe that this finding is inherent because students who have a lower level of high school education in mathematics, find it harder to master the material in technical/technological study programs.




(a) (b) (c) (d)

Fig. 4. Association between student dropout and: (a) gender; (b) study program; (c) enrolled 3rd semester; (d) categorical variables which indicates if the student passed at least two exams in the first year of study.

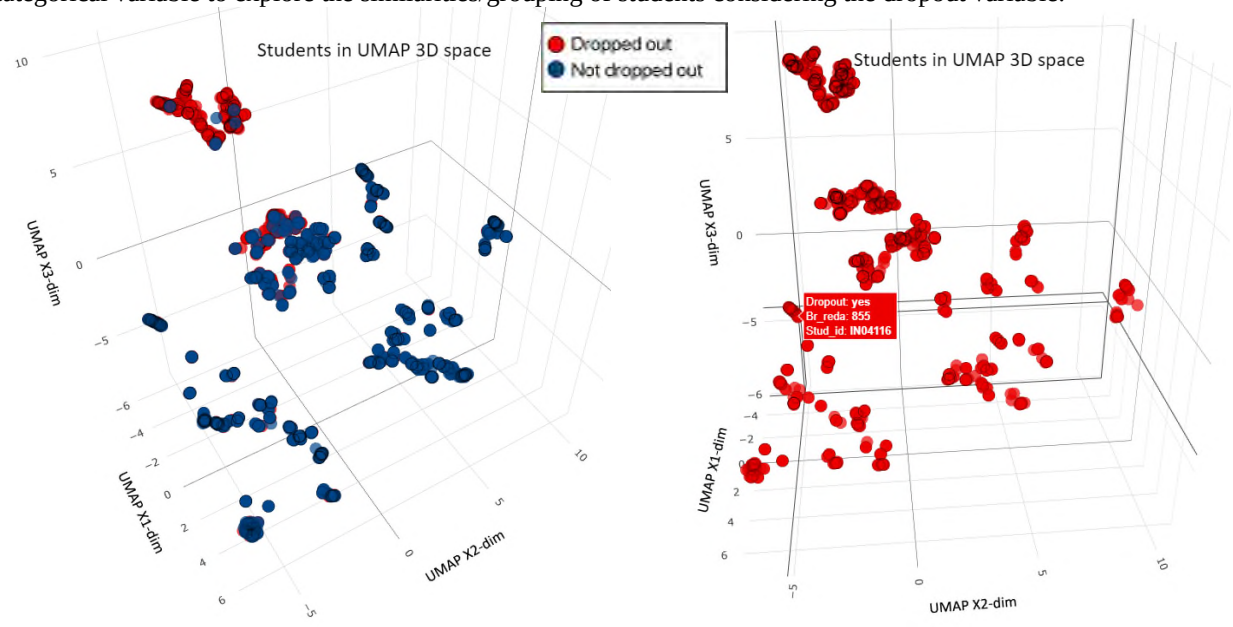
Bar plot representation of the categorical variable *enrolled3* show (Fig. 4c) that this variable is key for determining if the student will complete the studies or not. It significantly contributed to the reduction of potential variables. Similar conclusions can be drawn about the effects of following categorical variables on dropout: *first14* (which indicates if the student passed at least two exams in the first year of study) (Fig. 4d), *first37* (which indicates if the student achieved at least 37 ECTS in the first year of studies, and therefore fulfilled the condition for enrollment in the second year of studies) and *first48* (which indicates if the student achieved at least 37 ECTS in the first year of studies, and therefore fulfilled the condition for enrollment and state budget financing in the second year of studies). Bar plot representations of variables *first37* and *first38* isn't shown due to space limitations.

#### *3.2.2 Dimensionality reduction*

As mentioned in Section 2.2, UMAP, t-SNE, and FAMD dimensionality reduction is conducted to explore relationships between individuals by 3D visualization (Fig. 2, step B2).

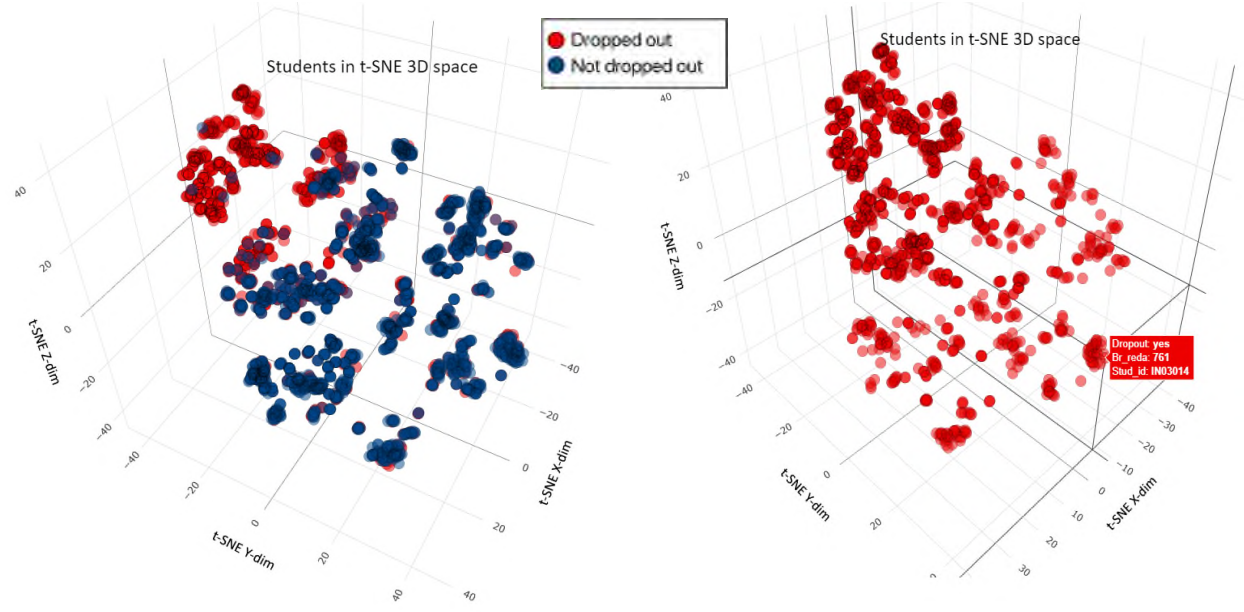
It can be seen (Fig. 5a) that students aren't separated well in UMAP three-dimensional space with respect to dropout, and there are multiple clusters where dropped out students are the majority class (Fig. 5b).

Fig. 6 shows the projection of students into first three t-SNE derived dimensions. Students are labeled with the dropout categorical variable to explore the similarities/grouping of students considering the dropout variable.




(a) (b)

Fig. 5. Students visualized in three UMAP derived dimensions: (a) All students are shown (b) Only dropped out students are shown.




(a) (b)

Fig. 6. Students visualized in three t-SNE derived dimensions: (a) All students are shown (b) Only dropped out students are shown.

It can be seen that students aren't separated well in t-SNE three-dimensional space (Fig. 6.a) with respect to dropout, and there are multiple clusters were dropped out students are the majority class (Fig. 6.b). Both reduction methods of dimension reduction indicate that the data is clusterable.

As pointed out before, FAMD is suitable for dimensionality reduction and exploration of mixed data type datasets (with both categorical and numeric variables). Also, this method is suitable for predictor importance analysis and exploring similarities between individuals by visualization in 2D/3D space.

Fig. 7 shows the projection of students into first three FAMD derived dimensions. Students are labeled with the dropout categorical variable to explore the similarities/grouping of students considering the dropout variable. It can be seen that the students are separated well in FAMD three-dimensional space with respect to dropout.

The visualization shows that the clustering of students is possible with regard to the dropout target variable, and that the students who dropped out are clearly distinguished in the space of first three FAMD dimensions.

Additionaly, Fig. 8 shows how much each variable contributed to FAMD dimensions. The larger the value, the bigger the contribution of a variable to a dimension. For example, the following variable contribute most to Dim.1: *y1_ects* (18.55), *y1_np* (18.14), *first37* (15.33), *first48* (13.55) and *first14* (13.04), and then *y1_mg* (9.94). The order of these contributions is sensible, given the meaning of the predictors.

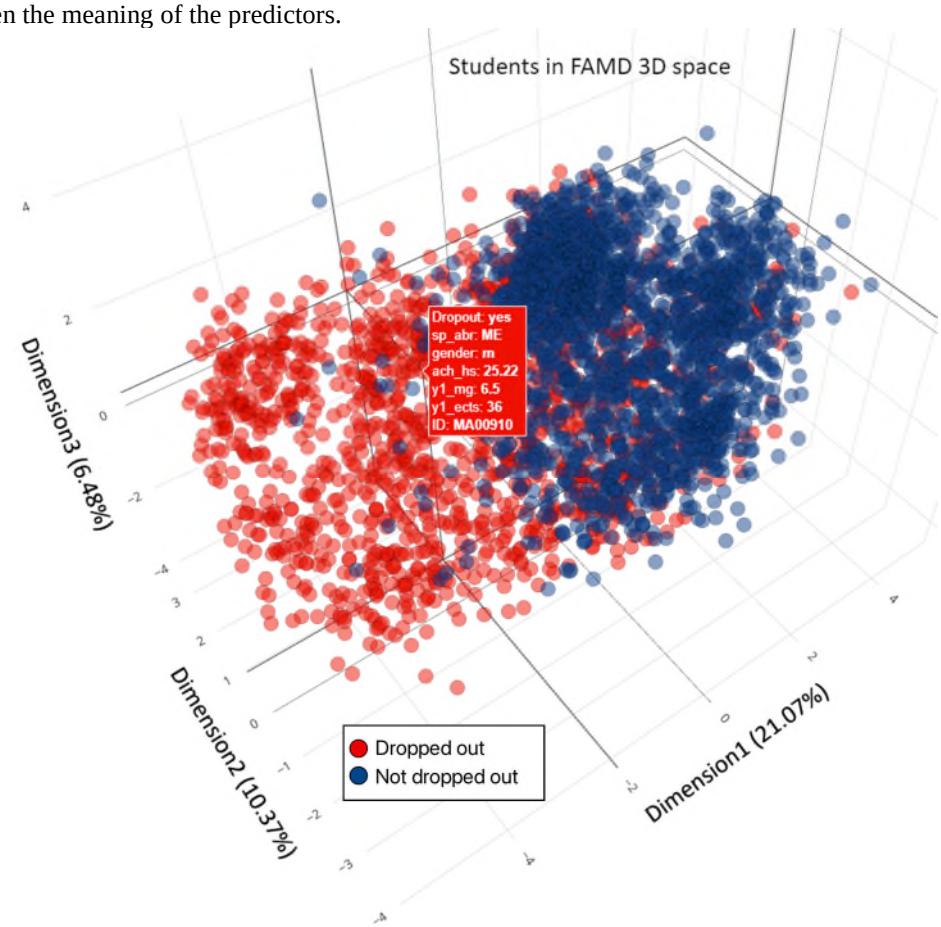



Fig. 7. Students visualized in three FAMD derived dimensions

For better interpreting FAMD dimensions, the representations given in Fig. 9 are suitable. These plots show the percentage contribution of variables from the original space to first three the dimensions in the new FAMD space. As an example of variable interpretation, the first two dimensions, Dim-1 and Dim-2, were selected. For example, the grade average in the first year of studies (*y1_mg*) contributes about 9%, and *gender* about 2% (Fig 9a) to Dim-1, while the achieved points from the general culture test at the entrance exam (*ach_ct*) contribute about 20% to Dim-2 (Fig. 9b).

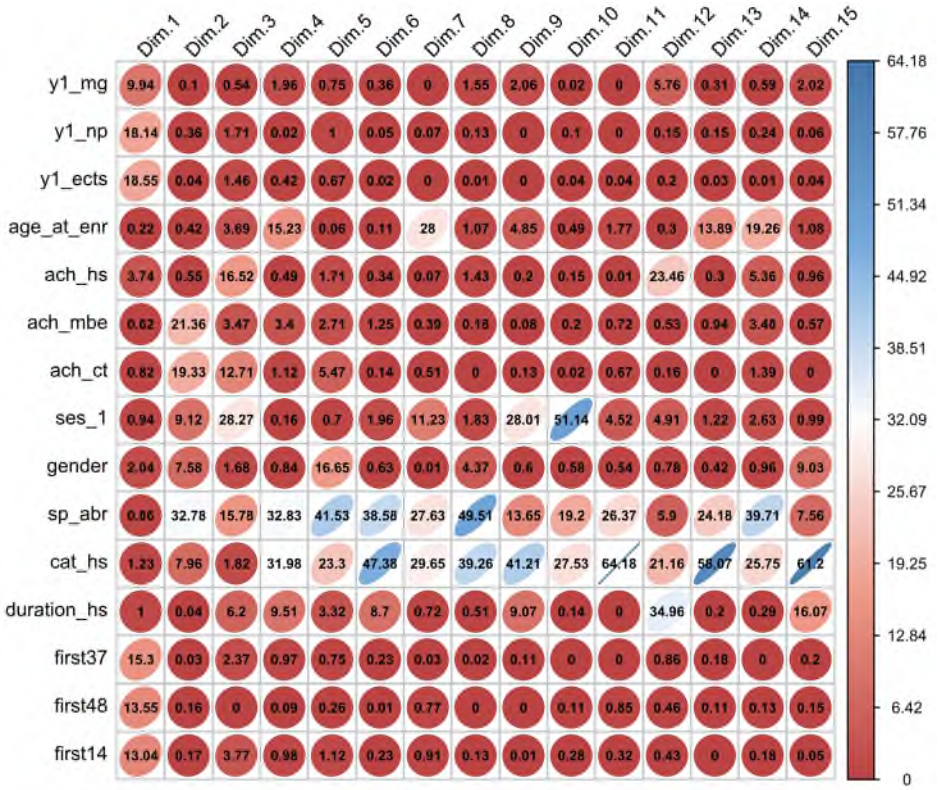
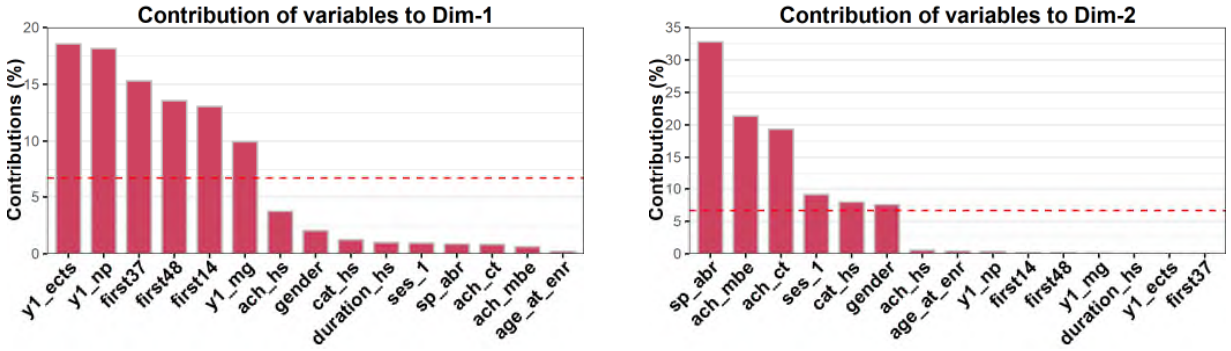



Fig. 8. Contribution matrix of variables to FAMD dimensions.




(a) Contribution of variables to Dimension 1

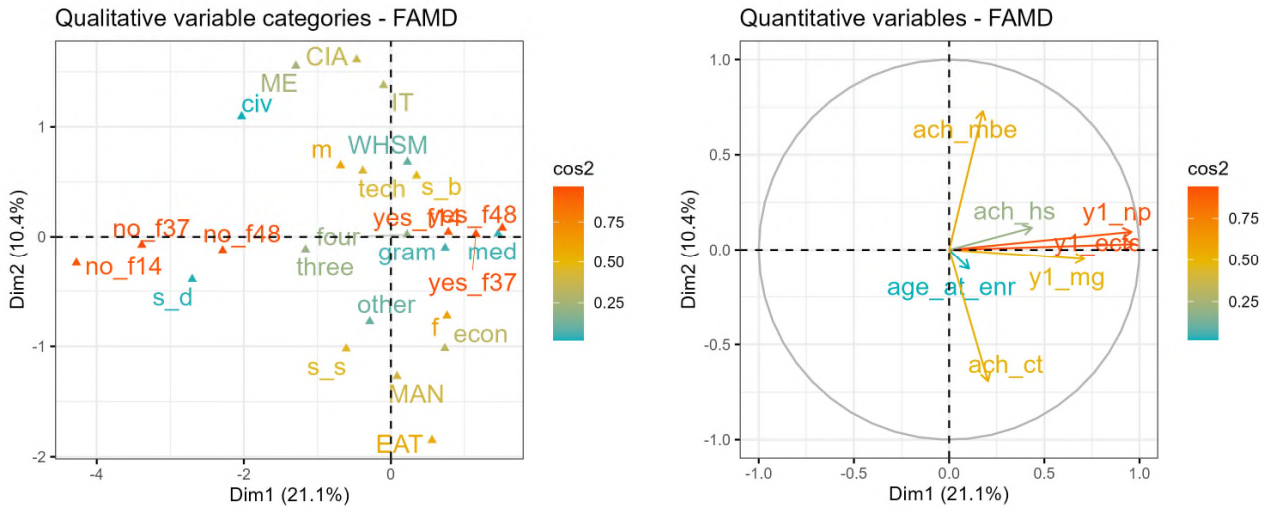
(b) Contribution of variables to Dimension 2

Fig. 9. Contribution of variables to first two FAMD dimensions

The dashed red line on the plot indicates the expected average contribution. For example, if the contribution of the variables were uniform, the expected value would be 1/(number of variables) = 1/15 = 6.66% (Fig 9.a)

Detailed analysis of qualitative variable categories and quantitative variables provides additional insights into quality of representation of numerical and categorical variables by levels (Fig10a) to each of the dimensions.

Variable categories which are similar are grouped together. Negatively correlated variables/variable categories are positioned in opposing quadrants. The distance between category points and the origin measures the quality of the variable category on the factor map. Category points that are away from the origin are well represented on the factor map.




(a) (b)

Fig. 10. cos2 - the quality of representation of the variables on FAMD factor map.

The quality of representation of the variables on factor map is called cos2 (squared cosine, squared coordinates). A high cos2 indicates a good representation of the variable in FAMD dimension. For example, categories IT and MAN of the *sp_abr* variable, are represented very well on factor map. Both categories, IT and MAN, are represented very well by Dim-2 and only negligibly by Dim-1, but with the opposite correlation sign (Fig10a).
Also, for the numerical variables, variables in a positive correlation are grouped together (Fig. 10b). Negatively correlated variables are placed in opposing quadrants. The closer a variable is to the circle of correlations, the better its representation is in Dim-1 and Dim-2 formed space (and it is more important to interpret them). If a variable is perfectly represented by only two FAMD dimensions (Dim.1 & Dim.2), the sum of the cos2 on these two dimensions is equal to one. In this case the variables will be positioned on the circle of correlations. For example, variable *y1_np* is one such variable. Vector *y1_np* (Fig. 10b), is almost of intensity 1, and the variable is dominantly represented by Dim-1 (projection onto Dim-1) and very negligibly by Dim-2 (projection onto Dim-2).

For an example of which set of variables we consider in this paper, the contribution of individual dimensions in reduced space, the explained variance is given in Fig. 11. It can be seen from the diagram that the reduction to the first three dimensions covers 37% (21.1 + 10.4 + 6.5) of the variance. The visualization shows that the clustering of students is possible with regard to the dropout target variable, and that the students who dropped out are clearly distinguished in the space of first three FAMD dimensions.

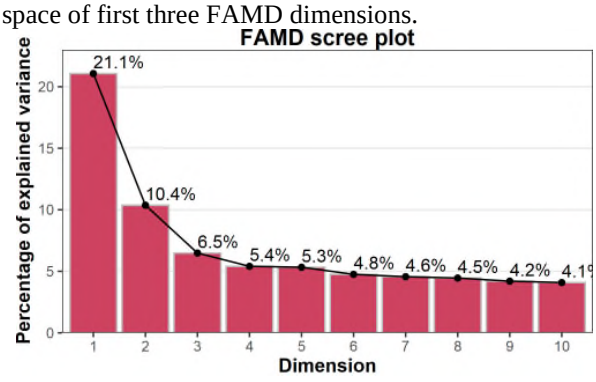



Fig. 11. Percentage contribution of FAMD dimensions to explained variance

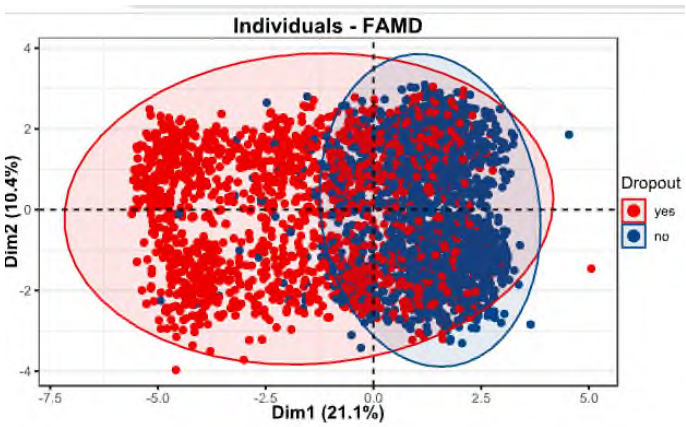



Fig. 12. Projection of students into 2D space of the first two FAMD derived dimensions

In addition to the depiction in Fig.7, Fig.12 also illustrates the projection of students into first two FAMD derived dimensions. Students are labeled with the dropout categorical variable to explore the similarities/grouping of students considering the dropout variable. It can also be seen that students are separated well in FAMD two-dimensional space with respect to dropout.

This analysis (based on steps B.1, B2., and domain knowledge) facilitated decision making in the final feature engineering phase (step B.3). Spurious columns/variables found highly correlated or redundant with other variables were identified and removed. A detailed description of selected features is shown is Table 2.

Table 2. Description of selected features kept for further analysis

| Feature | Type | Description | Value |
|---|---|---|---|
| *student_id* | nominal | Anonymized student ID | character |
| *y1_mg* | numeric | Mean grade on exams in the first 2 semesters | [5.0-10.0] |
| *y1_np* | numeric | Number of passed exams in the first two semesters | [0-10] |
| *y1_espb* | numeric | Achieved ECTS points in the first 2 semesters | [0-60] |
| *age_at_enr* | numeric | Age of student at enrollment date | [17-37] |
| *ach_hs* | numeric | Achieved points from high school education | [12-40] |
| *ach_mbe* | numeric | Points on the math/biology/economy test | [10-30] |
| *ach_ct* | numeric | Achieved points on the general culture test | [10-30] |
| *ses_1* | nominal | Student's funding status in the first 2 semesters | budget, self-financed, paused |
| *gender* | nominal | Gender of the student | male, female |
| *abr_sp* | nominal | Category name of the study program | CIA, IT, MA, MAN, EAT, WHSM |
| *cat_hs* | nominal | Secondary education category | tech, other, ec, gym, civ, med |
| *duration_hs* | nominal | Secondary education duration | three, four |
| *first37* | nominal | The student has 37 ECTS in first 2 semesters | TRUE, FALSE |
| *first48* | nominal | The student has 48 ECTS in first 2 semesters | TRUE, FALSE |
| *first14* | nominal | The student has 14 ECTS in first 2 semesters | TRUE, FALSE |
| *dropout* | nominal | Labeling student's dropout status | TRUE, FALSE |

Finally, after performing the above steps, the dataset based on selected variables is given as an input to train and test the clustering models.

### 3.3 Modeling

#### *3.3.1 Data splitting*

Based on feature engineering, after phase B, the resulting dataset is divided into Training and Testing datasets (Fig.2 step C.1). The Training dataset covered the period from 2007-2016 (3349 records), and the Test dataset students enrolled in 2017 (210 records).

#### 3.3.2 Clustering tendency, clustering validation, choice of the number of clusters

As described previously, clustering tendency assessment (Fig. 2, step C2) is performed by calculating the Hopkins statistic and inspecting visualizations produced in phase B for meaningful groupings. A value of 0.1054109 was calculated for the H statistic which is significantly less than the threshold value of 0.5. This is further confirmation that the dataset is clusterable.

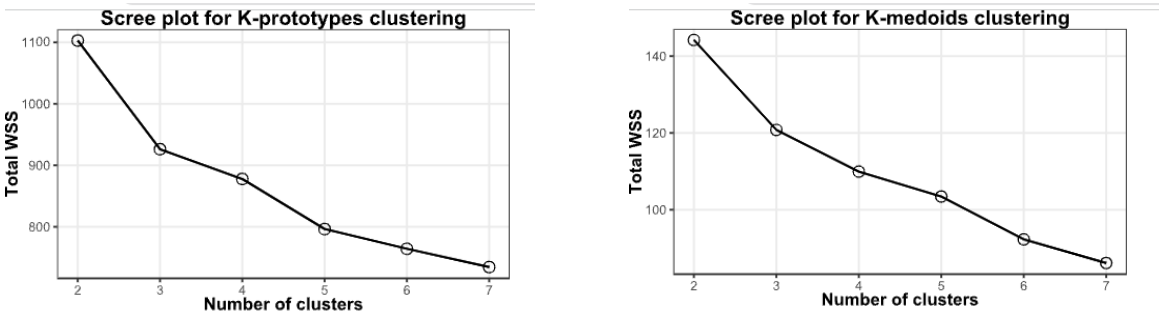
The optimal number of clusters was estimated based on a large number of metrics for internal clustering validation. (details are given in Section 2.2). The calculation of validation metrics was carried out for clustering models with number of clusters ranging from two to seven. Due to size limitations, the values for these metrics are shown only for models with 2, 3, and 4 clusters.

Additionally, for hierarchical clustering best linkage method was chosen based on correlation between the *cophenetic distances* [34], and the Gower's dissimilarity matrix [35] of the training dataset, which was used to cluster the data by hierarchical clustering. The Mcquitty linkage method [36] was chosen based on the highest correlation value between cophenetic distances and distances between students in the Gower distance matrix — 0.742267, which is in line with recommendations made in [ref].

For all metrics, except entropy, the optimal number of clusters corresponds to the maximum value of the metric. The weighted majority voting rule between the validation metrics was used to decide on the final number of clusters to be used. The Silhouette method was weighted more than other metrics — and was given two votes. It can be seen in Table 3. That the dominant choice for the number of clusters is 2 or 3 clusters (Fig. 2, step C2). This is also confirmed by the Scree plots for k-medoids and k-prototypes clustering models (Fig. 13), where the first significant slope change for WSS is at the point denoting the third cluster.

Table 3. Internal metrics for clustering validation

| Clustering method | Number of clusters | Sillouete width | Within_ clust_ss | Dunn Index | Dunn2 Index | Calinski Harabasz Index | Wb_ratio Average within/between distance ratio | Entropy |
|---|---|---|---|---|---|---|---|---|
| AGL_HC | 2 | 0.386 | 139.531 | 0.125 | 1.574 | 1992.026 | 0.604 | 0.567 |
| AGL_HC | 3 | 0.276 | 134.119 | 0.122 | 1.172 | 1103.409 | 0.599 | 0.684 |
| AGL_HC | 4 | 0.261 | 130.097 | 0.110 | 1.024 | 792.587 | 0.599 | 0.828 |
| PAM | 2 | 0.359 | 144.205 | 0.010 | 1.498 | 1818.972 | 0.628 | 0.583 |
| PAM | 3 | 0.233 | 120.784 | 0.019 | 1.114 | 1409.920 | 0.645 | 1.062 |
| PAM | 4 | 0.199 | 109.932 | 0.017 | 1.044 | 1142.506 | 0.637 | 1.380 |
| Two-step | 2 | 0.363 | 141.265 | 0.012 | 1.455 | 1926.501 | 0.625 | 0.606 |
| Two-step | 3 | 0.195 | 125.147 | 0.019 | 0.950 | 1302.445 | 0.663 | 1.082 |
| Two-step | 4 | 0.184 | 110.792 | 0.022 | 1.050 | 1124.972 | 0.629 | 1.288 |
| gower | | | | | | | | |
| | | Silhouette | | | | | | Mcclain |
| k-Prototypes | 2 | 0.261 | | | | | | 0.723 |
| k-Prototypes | 3 | 0.317 | | | | | | 0.549 |
| k-Prototypes | 4 | 0.272 | | | | | | 0.529 |




(a) (b)

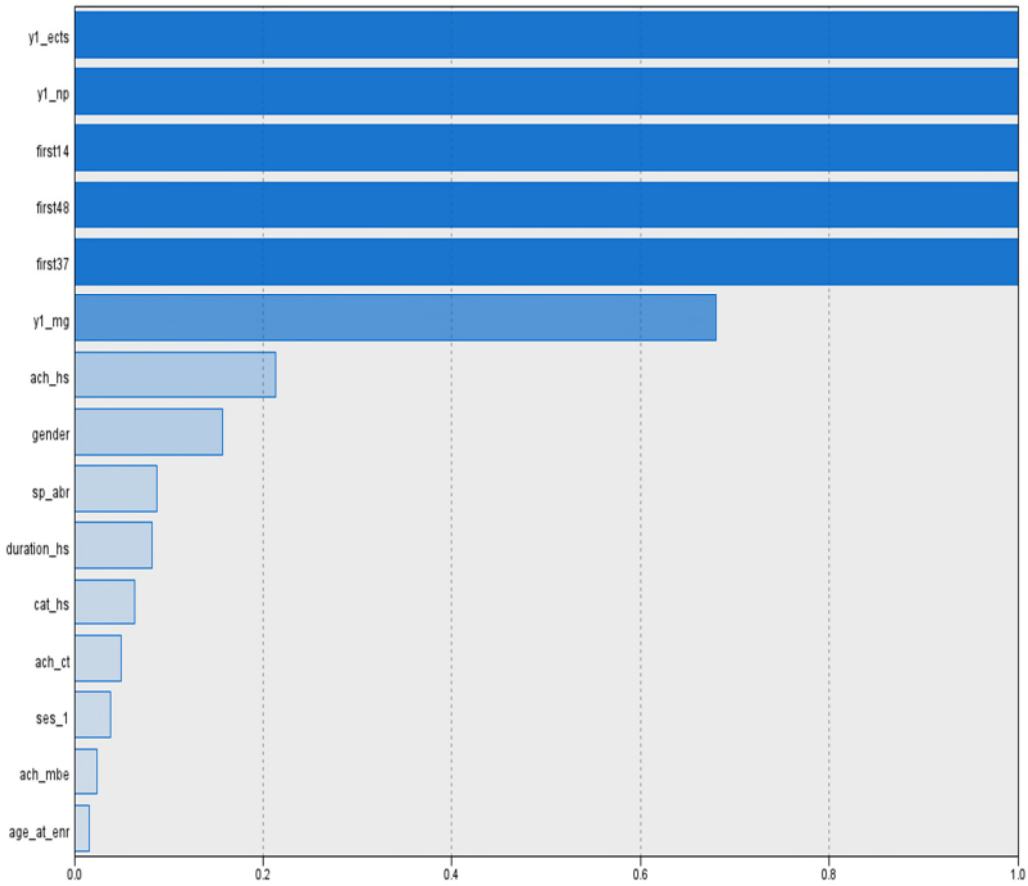
Figure 13. Scree plots for K-prototypes (a) and K-medoids (b) clustering

### 3.3.3 Variable importance analysis for clustering

Two-step clustering also calculates variable importance for input variables (Fig. 2, step C5a) based on the algorithm described in [22]. The importance $V$ of variable $i$ is defined as:

$$VI_i = \frac{-log_{10}(sig_i)}{max_{j\in\Omega}(-log_{10}(sig_i))}$$

where $\Omega$ denotes the set of predictor and evaluation field, $sig_i$, is the significance or $p$ - value computed rom applying a certain test. If $sig_i$ equals zero, set $sig_i = MinDouble$ , where $MinDouble$is the minimal double value and where $j$ is the $j$ -th cluster number. The more important the variable the close the $V$ value is to 1. Variable importances for the two-step clustering model for two clusters, are given in Fig. 14.




| Nodes | Imp. |
|---|---|
| first37 | 1 |
| first48 | 1 |
| first14 | 1 |
| y1_np | 1 |
| y1_ects | 1 |
| y1_mg | 0.680 |
| ach_hs | 0.213 |
| gender | 0.157 |
| sp_abr | 0.088 |
| duration_hs | 0.083 |
| cat_hs | 0.064 |
| ach_ct | 0.050 |
| ses_1 | 0.039 |
| ach_mbe | 0.024 |
| age_at_enr | 0.016 |

Figure 14. Variable importance based on Two-step clustering method (SPSS…)

Based on the obtained values the variables can be classified into three categories according to their importance. The biggest effect comes from the variables that indicate: the number of achieved ECTS (*y1_ETCS*), the number of passed exams in the first year of studies (*y1_np*), as well as binary variables that indicate whether the student has achieved the threshold number of points for enrollment in the next year of studies (*first37*), if the student is budget-financed (*first48*), if the student passed at least two exams in the first year of studies (*first14*), and mean his grade (*y1_mg*). Thus, the primary factors are related to learning achievement.

The next group of variables consists of: achieved points during secondary education (*ach_hs*), *gender*, duration of secondary education (*duration _hs*), category of secondary education (technical, economics, grammar school…) (*cat_hs*), as well as the study program he enrolled in (*sp_abv*). The variable importances for the first and second group of variables is expected.

Finally, the least influential group of factors are: age at enrollment (*age_at_enr*), number of points in mathematics and general knowledge entry test (*ach_mbe*, *ach_ct*). The fact that success in entry tests for mathematics, biology, or economics (as key to appropriate study programs) shows little importance, has no logical basis in our opinion. So, these variables require additional analyses.

Variable importances for the two-step clustering model for three clusters are the same as in the two clusters models, so they aren't shown here.

### 3.3.4. Clustering

Clustering on the training dataset (Fig. 2, step C5) is performed for each clustering method (k-prototypes, two-step, PAM, and hierarchical clustering) and a selected number of clusters (K=2, or K=3) for those methods. K-medoids clustering, and Hierarchical agglomerative aren't formed on the training dataset, but are instead formed on the basis of the Gower distance matrix of the training dataset. As a results, segmentation of students was achieved by multiple clustering models.

### 3.4 Evaluation

Classification of students with regards to the *dropout* variable was performed in phase D (Fig. 2, step D1, and D2).

In order to perform classification of students it was first necessary to form classification models based on the formed clustering models. This was accomplished as follows. For each clustering models, the cluster with most dropped out students is chosen as a representative "dropout cluster" for that model. Then, cluster labels with that cluster number serve as "dropped out" labels, and all other cluster labels serve as "not dropped out" labels. With this scheme, confusion matrices

and metrics (Table 4) — precision, recall, specificity, accuracy, balanced accuracy — are produced for both the test and training dataset, for all cluster models.

Table 4. Description of used evaluation metrics

| | | Evaluation metric | Description |
|---|---|---|---|
| | | $Precision = \frac{TP}{TP + FP}$ | Precision shows what proportion of positive predictions was actually correct. Precision is a more appropriate metric when it is important to have as few false positive classifications as possible. A model that produces no false positives has a precision of 1.0. |
| | | $Recall = \frac{TP}{TP + FN}$ | Recall (sensitivity, true positive rate) shows what proportion of actual positives was predicted correctly. Recall is a more appropriate metric when it is important to have as few false negative classifications as possible. A model that produces no false negatives has a precision of 1.0. |
| | | $Specificity = \frac{TN}{TN + FP}$ | Specificity (true negative rate) shows what proportion of actual negatives was predicted correctly. A model that produces no false positives has a precision of 1.0. |
| | | $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ | Accuracy shows the proportion of correct predictions. |
| | | $\text{Balanced Accuracy} = \frac{Recall + Specificity}{2}$ | Balanced accuracy is the arithmetic mean of sensitivity and specificity, and it is appropriate when dealing with imbalanced data. |

| | | Predicted dropout | |
|---|---|---|---|
| | | NO | YES |
| Actual dropout | NO | **TN** | **FP** |
| | YES | **FN** | **TP** |

The values of the obtained metrics, which describe the quality/performance of the obtained models, are often presented in the form of confusion matrices. For example, Fig. 15 shows the confusion matrix for the clustering/classification model obtained by the two-step clustering method based on three clusters.

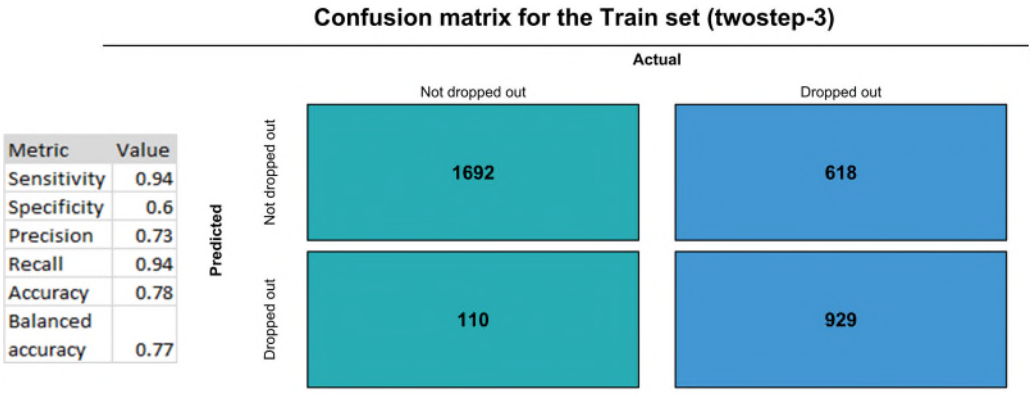



Figure 15. Confusion matrix for the predictions made by the Two-step based clustering model with three clusters on the training dataset

For easier comparison and analysis, classification metrics values for all clustering models are given in Table 5.

Analysis of the values of the *recall* metrics listed in Table 5 shows: (1) Agglomerative hierarchical clustering on three clusters does not make sense; (2) Although the recall metric has similar values for all clustering models with three clusters on the training dataset (0.98; 0.94; 0.98), the models based on PAM and k-protypes clustering have superior performance on the test dataset; (3) From the group of clustering models with two clusters, the two-step based model has worst performance, and classification based on this model does not make sense.

PAM (0.96 on the Train dataset and 1.00 on the Test dataset) and AGL_HC (0.98 on the Train dataset and 1.00 on the Test dataset) have highest recall values, 1.0. This means that the PAM and AGL_HC models produce no false negative on the test dataset, i.e., they identify 100% of the students that actually dropped out in the test dataset.

Table 5. Classification metrics for all clustering models

| | Number of clusters | Subset | Precision | Recall | Specificity | Accuracy | Balanced Accuracy |
|---|---|---|---|---|---|---|---|
| k-prototypes | 3 | Train | ***0.68*** | 0.98 | 0.47 | 0.75 | 0.73 |
| | 3 | Test | ***0.57*** | **1.00** | 0.39 | 0.66 | 0.69 |
| | 2 | Train | 0.73 | 0.89 | 0.62 | 0.77 | 0.76 |
| | 2 | Test | 0.58 | 1.00 | 0.42 | 0.68 | 0.71 |
| Two-Step | 3 | Train | 0.73 | 0.94 | 0.6 | 0.78 | 0.77 |
| | 3 | Test | 0.39 | 0.57 | 0.27 | 0.40 | 0.42 |
| | 2 | Train | 0.72 | 0.95 | 0.58 | 0.78 | 0.76 |
| | 2 | Test | 0.00 | 0.00 | 0.46 | 0.25 | 0.23 |
| PAM | 3 | Train | ***0.69*** | 0.98 | 0.49 | 0.75 | 0.73 |
| | 3 | Test | ***0.59*** | **1.00** | 0.43 | **0.69** | 0.72 |
| | 2 | Train | 0.71 | *0.96* | 0.54 | 0.76 | 0.75 |
| | 2 | Test | 0.60 | 1.00 | 0.46 | **0.70** | 0.73 |
| AGL_HC | 3 | Train | 0.04 | 0.02 | 0.47 | 0.23 | 0.25 |
| | 3 | Test | 0.00 | 0.00 | 0.56 | 0.31 | 0.28 |
| | 2 | Train | 0.71 | *0.98* | 0.53 | 0.77 | 0.75 |
| | 2 | Test | 0.59 | 1.00 | 0.44 | 0.69 | 0.72 |

The *precision* metric has worse values for all models compared to *recall*. From the group of clustering models with three clusters, AGL_HC and Two-step models have the worst performance w.r.t the precision metric. The AGL_HC model with three clusters has terrible performance even for the train dataset. Other three cluster-based models, On this basis, k-prototypes and PAM, methods have similar performance (≈0.68; ≈0.57) w.r.t the recall metric. From the group of models based on two clusters, Two-step model has the worst performance, and the three remaining models have similar performance (≈0.71 on the Train dataset and ≈0.59 on the Test dataset). It can be seen that models based on PAM and k-prototypes have Precision ≈0.58 regardless of the number of clusters the model is based on.
Accuracy is roughly ≈0.75 for the training dataset on all models, and roughly ≈0.70 on the test dataset for models k-prototypes and PAM.
If the classification metrics for the test are not satisfactory, corrections are made in step B.3, and/or step C.2.

## 4 DISCUSSION

The results of the implementation of the proposed methodology through a series of steps are explained in detail and discussed. Therefore, only selected final observations are given related to the creation of a student dropout prediction model.
Although the data preparation and EDA phases relied on complete data, i.e., data that contains predictors related to the second and third year of studies, it turned out that the variable that denotes enrollment in the third semester is the most powerful predictor, and makes a clear distinction between those who graduated and those who left their studies. Based on this fact, the authors removed a large part of the variables that pertain to the 2nd and 3rd year of studies, thus significantly reducing the set of predictors.
Based on the analysis realized in phase C (modeling) and the comparative analysis described in section 3.4, the PAM model we chose the PAM model as the optimal prediction model for student dropout.
This model is characterized by approximately similar performance regardless of whether the number of clusters is K = 2 or K = 3 (Fig.16).

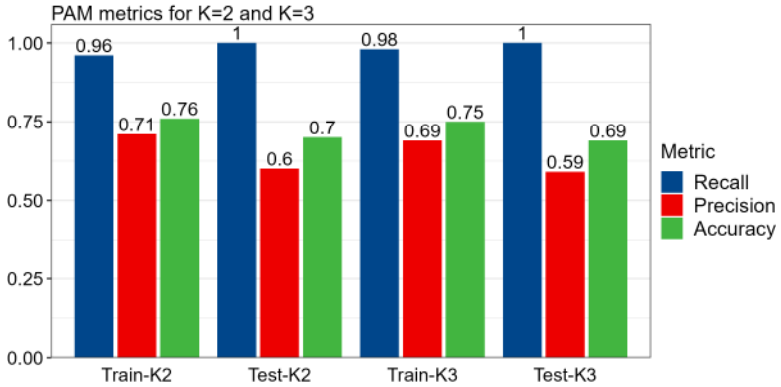
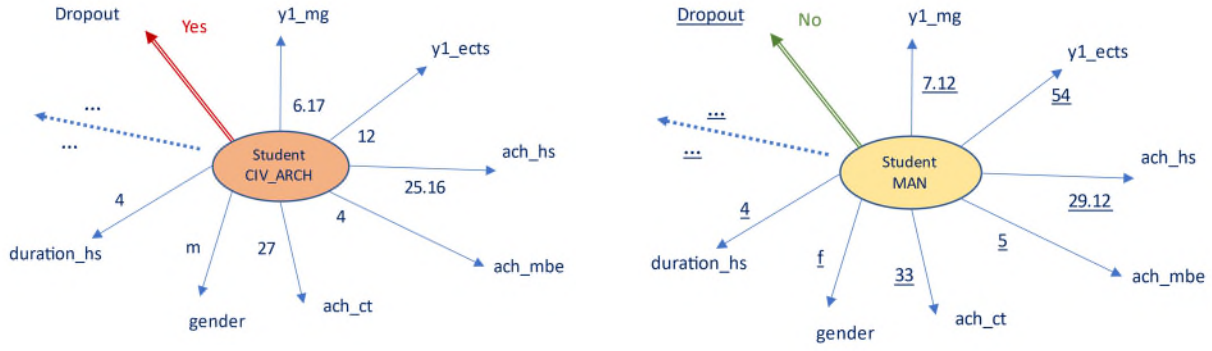


Fig. 16 Classification metrics for PAM based models

For the case K = 2, the representative of the dropout cluster is a civil engineering student with a very low mean grade in the first year of studies (*y1_mg* = 6.00 out of 10), who achieved only 12 *y1_ects* out of 60, passed only two exams, finished a technical high school, had medium high school performance (*ach_hs* = 25.16 out of 40), achieved only 4 points out of

30 on the mathematics entry exam, and he is male (Fig. 17a). In a similar way, a medoid for the second cluster is a representative of students who have completed their studies (Fig. 17b).
The division into three clusters achieves a division into finer groups. The first group, the dropout group, is similar to the dropout group when K = 2. And the second (students of work health and safety engineering) and the third (student of management) group is represented by similar medoids (one is a student of engineering) primarily distinguished by the results achieved on the mathematics entrance exam.




(a) (b)

Fig. 17 PAM medoids for K=2: (a) Medoid for Dropout students and (b) Medoid for Non-Dropout students

## 5 CONCLUSION

Policies related to student retention are becoming increasingly important in countries in Europe and the around world, as the effects of student dropout have far-reaching implications for the development perspective of the knowledge economy. In the Republic of Serbia, the number of students at the Academies of Vocational Studies has been declining in the last 5-10 years, so the issue of student dropout for these institutions is of immense importance, both in terms of income and alumni contributions and in terms of public image of the institution itself.
For these reasons, this study focuses on conceptualizing a methodology for creating prediction models of student dropout. The methodology is based on clustering models which are then adapted into classification methods, and the hypothesis was that it is possible to develop a model with acceptable performance for early detection of student dropout.
To the best of our knowledge, this is the first study in the Republic of Serbia that uses machine learning as a platform for modeling student dropouts.
Models were developed by using the database provided by the Student service of the Western Serbia Academy of Applied Studies, Uzice, which contained data the for the period of 2007-2020 for over 3000 students. Data preparation and exploratory data analysis was realized by using a wide range of statistical methods and dimensionality reduction algorithms. The contribution of predictors to selected derived dimensions was analyzed in detail. Comprehensive feature engineering, which included Variable Importance analysis, resulted in a set of variables which are relevant for the student dropout phenomenon. The variables are divided into 4 groups: (1) demographic variables, (2) variables describing secondary education achievement and high schools (3) variables that denote student's achievement in the entry exams, and finally (4) student's achievements during the first year of studies. Based on a large number of clustering validation metrics the authors decided to segment students into two or three clusters. For the selected binary variable *dropout* (that indicated if the students dropped out after 4 years of study) the following prediction models were evaluated on the test dataset: *k-prototypes, k-medoids* (PAM), *two-step clustering*, and *agglomerative hierarchical clustering*.
PAM was selected as the best model for detection of student dropout based on its performance on the test dataset, evaluated by classification metrics: Recall=1, Precision=0.70, and Accuracy= 0.75. This confirmed the initial hypothesis of this study.
In order to improve the methodology developed in this study, and the performance of the prediction models in future research, the authors plan to: (1) Perform detailed analysis of additional variables that could contribute to early detection of students who tend to leave their studies. Examples of these variables are: student's pre-examination achievements on colloquiums, seminar work, and homework for the courses in the first semester, education level of the student's parents, etc. (2) enrich and improve prediction models in terms of interpretability and performance by using tree-based classification algorithms such as: Decision Trees, Random Forests, Gradient Boosting Machines, etc.
The authors believe that models of early detection of student dropouts, based on Machine learning, can significantly contribute to proper decisions and development policies of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

## REFERENCES

[1] S. Wild, L.S. Heuling, Student dropout and retention: An event history analysis among students in cooperative higher education, International Journal of Educational Research. 104 (2020) 101687.
[2] B.M. Kehm, M.R. Larsen, H.B. Sommersel, Student dropout from universities in Europe: A review of empirical literature, (2019).
[3] T.Z. Zając, A. Komendant-Brodowska, Premeditated, dismissed and disenchanted: higher education dropouts in Poland, Tertiary Education and Management. 25 (2019) 1–16.

[4] M.S. DeBerard, G.I. Spielmans, D.L. Julka, Predictors of academic achievement and retention among college freshmen: A longitudinal study, College Student Journal. 38 (2004) 66–81.
[5] D. Delen, Predicting student attrition with data mining methods, Journal of College Student Retention: Research, Theory & Practice. 13 (2011) 17–35.
[6] OECD, S.O. of the E. Communities, Oslo Manual, (2005). https://www.oecd-ilibrary.org/content/publication/9789264013100-en.
[7] B. Prenkaj, G. Stilo, L. Madeddu, Challenges and solutions to the student dropout prediction problem in online courses, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020: pp. 3513–3514.
[8] STATISTICAL OFFICE OF THE REPUBLIC OF SERBIA, MUNICIPALITIES AND REGIONS OF THE REPUBLIC OF SERBIA, 2011, (2011). https://publikacije.stat.gov.rs/G2015/PdfE/G20152017.pdf.
[9] STATISTICAL OFFICE OF THE REPUBLIC OF SERBIA, STATISTICAL RELEASE AS20, 149 (2020) A33–A34. https://doi.org/10.1016/s0022-5223(14)01847-9.
[10] L. Aulck, N. Velagapudi, J. Blumenstock, J. West, Predicting student dropout in higher education, ArXiv Preprint ArXiv:1606.06364. (2016).
[11] L. Kemper, G. Vorhoff, B.U. Wigger, Predicting student dropout: A machine learning approach, European Journal of Higher Education. 10 (2020) 28–47.
[12] S. Supangat, M.Z.B. Saringat, G. Kusnanto, A. Andrianto, Churn Prediction on Higher Education Data with Fuzzy Logic Algorithm, SISFORMA. 8 (2021) 22–29.
[13] T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.C. Chatzisavvas, A comparison of machine learning techniques for customer churn prediction, Simulation Modelling Practice and Theory. 55 (2015) 1–9.
[14] O. Çelik, U.O. Osmanoglu, Comparing to techniques used in customer churn analysis, Journal of Multidisciplinary Developments. 4 (2019) 30–38.
[15] K. Palani, P. Stynes, P. Pathak, Clustering Techniques to Identify Low-engagement Student Levels., in: CSEDU (2), 2021: pp. 248–257.
[16] Á.A.M. Navarro, P.M. Ger, Comparison of clustering algorithms for learning analytics with educational datasets, IJIMAI. 5 (2018) 9–16.
[17] N. Iam-On, T. Boongoen, Generating descriptive model for student dropout: a review of clustering approach, Human-Centric Computing and Information Sciences. 7 (2017) 1–24.
[18] M. Macedo, C. Santana, H. Siqueira, R.L. Rodrigues, J.L.C. Ramos, J.C.S. Silva, A.M.A. Maciel, C.J. Bastos-Filho, Investigation of college dropout with the fuzzy c-means algorithm, in: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), IEEE, 2019: pp. 187–189.
[19] V. Hegde, P. Prageeth, Higher education student dropout prediction and analysis through educational data mining, in: 2018 2nd International Conference on Inventive Systems and Control (ICISC), IEEE, 2018: pp. 694–699.
[20] I. Stepanović-Ilić, O. Tošković, K. Krstić, M. Videnović, Dropout at university level in Serbia: Analysis of measurement, research findings, services and prevention measures, Zbornik Instituta Za Pedagoska Istrazivanja. 52 (2020) 479–519.
[21] S.M. Cisar, R. Pinter, Analysis of students' dropout rate at Subotica Tech, Journal of Applied Technical and Educational Sciences. 9 (2019) 43–54.
[22] IBM Corp, IBM SPSS Statistics Algorithms, (n.d.) 165,166.
[23] R. Wirth, J. Hipp, CRISP-DM: Towards a standard process model for data mining, in: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Springer-Verlag London, UK, 2000.
[24] D.B. Rubin, Inference and missing data, Biometrika. 63 (1976) 581–592.
[25] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art., Psychological Methods. 7 (2002) 147.
[26] R.J. Little, A test of missing completely at random for multivariate data with missing values, Journal of the American Statistical Association. 83 (1988) 1198–1202.
[27] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, ArXiv Preprint ArXiv:1802.03426. (2018).
[28] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE., Journal of Machine Learning Research. 9 (2008).
[29] J. Pagès, Analyse factorielle de données mixtes, Revue de Statistique Appliquée. 52 (2004) 93–111.
[30] A. Banerjee, R.N. Dave, Validating clusters using the Hopkins statistic, in: 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542), IEEE, 2004: pp. 149–153.
[31] C. Hennig, Flexible procedures for clustering, Version 2.1-11.1. (2018).
[32] R. Aschenbruck, G. Szepannek, Cluster Validation for Mixed-Type Data, Archives of Data Science, Series A. 6 (2020) 02.
[33] H. Wickham, The tidyverse, R Package Ver. 1 (2017) 836.
[34] S. Saraçli, N. Doğan, İ. Doğan, Comparison of hierarchical cluster analysis methods by cophenetic correlation, Journal of Inequalities and Applications. 2013 (2013) 1–8.
[35] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics. (1971) 857–871.
[36] L.L. McQuitty, Similarity analysis by reciprocal pairs for discrete and continuous data, Educational and Psychological Measurement. 26 (1966) 825–831.