Future Generation Computer Systems 125 (2021) 661–671

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Student success prediction using student exam behaviour

Jakub Kuzilek [a,b,*], Zdenek Zdrahal [a], Viktor Fuglik [a,c]

[a] CTU in Prague, CIIRC, Jugoslavskych partyzanu 1580/3, Prague, Czech Republic
[b] Humboldt University of Berlin, Unter den Linden 6, Berlin, Germany
[c] Charles University, Computer Science Centre, Ovocny trh 560/5, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 6 October 2020
Received in revised form 23 June 2021
Accepted 7 July 2021
Available online 14 July 2021

Keywords:
Student dropout
Predictive modelling
Exam behaviour
Behavioural patterns

ABSTRACT

The Faculty of Mechanical Engineering, Czech Technical University in Prague (FME) faces a significant student drop-out in the first-year bachelor programme, which is an actual problem for many higher education institutions. Metacognitive processes play a vital role in self-regulated learning. Students become active participants in their learning, and one critical aspect of higher education studies is planning and time management. The exam taking behaviour is in the context of the FME manifestation of the time management skills of each student; thus, the exam-taking patterns may help identify at-risk students. To evaluate the importance of exam behaviour patterns, we conducted three experiments. Identification of students passing or failing the first study year has been conducted using four different machine learning models. The exam taking behaviour patterns increase the prediction F-measure significantly for the class of failing students (approximately 0.3 increase). Moreover, the approach based on student behaviour enabled us to identify the critical exam-taking patterns, which further helps the lecturers identify at-risk students and improve their time management skills and chances to pass the first academic year.

1. Introduction

Recent years have brought a new era of digital technologies to higher education. The massive employment of Information and Communication Technologies (ICT) [1] in delivering the learning content to the students of typical higher education institutions results in the collection of an increasing quantity of study- and student-related data. These data are in various forms and includes, for example, student demographics, course contents, student results, etc. Nowadays, between 20% and 54% of students fail to complete their degrees [2]. This trend is especially grave in the engineering disciplines [3,4]. Introductory science courses are the most challenging, with high failure rates leading to discouraging students from pursuing Science, Technology, Engineering and Mathematics degrees [5].

For the first-year (newcomers) student's success, multiple critical aspects have been identified, including instructional styles, faculty expectations or the learners' cognitive, behavioural, motivational and developmental capabilities [6]. To succeed academically, first-year students need to adapt to the new university life and learn how to study and efficiently manage their time and optimize their study practices [7]. Thus, educational researchers interested in metacognition and self-regulated learning (SLR) [8]. SLR is the process in which students act as active participants in their learning. Mainly, three aspects of SRL manifest in learners who are proactive and systematic [7]: metacognitive processes (planning, goal setting, monitoring the learning progress, self-evaluating); motivation processes (interest in studies, accepting responsibility) and behavioural processes (seeking information and advice, adopting effective study strategies). When focusing on metacognitive processes, students recognize the goal settings and planning/time management very important skills to master to succeed in their academic career [7]. To address the challenges students face during their studies, many institutions seek new student support possibilities. One promising direction is using the student data to understand educational processes better and identify the students at risk [9].

Our research focuses on traditional brick and mortar university in the Czech Republic, which follows the typical educational scheme in post-Austrian-Hungary countries [10]. University students study several courses during the semester. The semester is followed by the exam period. To successfully progress through the exam period and navigate through their defined study trajectory based on a study plan, students need to learn how to plan and manage their exams within the predefined time. Thus the students need to acquire at least some knowledge of how to organize efficiently their time (they must "learn how to learn"). In this regard, even when the university collects only data regarding

[*] Corresponding author at: Humboldt University of Berlin, Unter den Linden 6, Berlin, Germany.
E-mail address: jakub.kuzilek@gmail.com (J. Kuzilek).

https://doi.org/10.1016/j.future.2021.07.009
0167-739X/© 2021 Elsevier B.V. All rights reserved.

the study progression, such as course results and does not focus on learning itself, students at risk of failure can be identified [11].

The rest of the section will introduce related work, research goals, and the educational setting our research explores. The rest of the paper then covers the data and methodology used, the methods employed for the analysis and finally, our conclusions.

1.1. State of the art

The boom of ICT brings new possibilities for measuring, storing, and analysing various processes and phenomena. Higher education is no exception, and universities use different tools to collect and analyse study-related data [9]. For the analysis of data collected from university (study) information systems, including Virtual Learning Environments (VLEs), which has been proved to be the source of helpful information [12,13], and various analytical techniques have been investigated.

Baradwaj and Pal [14] deployed a decision tree to predict the performance of 50 students using course and study-related attributes to identify students with additional support required.

Shehata and Arnold [15] reported the deployment of a so-called "regression predictive engine" for predicting students' performance at the University of Wisconsin using the social, demographical and educational features to identify the students at risk of failing their studies.

Wolff et al. [16–18] deployed multiple predictive models including *k*-Nearest Neighbours, decision trees or Naive Bayes classifier for identification of students at risk of failing based on student demographical, performance-based and online learning trace data. The identified students are then marked for additional support provided by the lecturers of the university.

Xenos [19] reported an approach for modelling student behaviour using the Bayesian network to identify students' improper habits and identify students at risk of failing their degrees.

Howard et al. [20] developed an approach for predicting the final course marks based on Bayesian Additive Regressive Trees. The method can identify at-risk students by week 6 of the semester using online learning trace data.

In [21] Huang and Fang built the models for identifying students' success or failure using student previous study achievements. It has been observed that predicting failure within first-year courses is critical because the failure rate is usually high, yet many students can be saved with additional support.

The student demographics, combined with information about their study history without any VLE related data, can also be used to estimate study success [22]. If no demographical information is available, prediction models can still use the student's past performance (grades) [23].

One of the fundamental ideas presented in all current research is using legacy data to develop predictive models and apply these models to make predictions on currently running courses. This approach is helpful, especially for the course lecturers and staff responsible for implementing interventions to improve student retention rate and students themselves.

Student performance modelling can be further extended to the analysis of student activity (behaviour). Hlosta et al. [24] proposed two methods for student activity analysis in VLEs: General Unary Hypothesis Automaton and Markov chains. The idea was extended by Okubo et al. [25]. More recent work of Davis et al. [26] employed Markov chains to analyse MOOC data from edX and Coursera courses with over 100,000 students.

Furthermore, machine learning methods can be used in the analysis of face-to-face learning. For example, Kent et al. [27] explored the relationship between student engagement and learning outcomes in the context of traditional brick and mortar UK university. They examined multiple university data sources containing data about student demographics, performance and the use of library resources. Their study was performed on more than 30,000 students, and researchers have predicted aggregative student scores.

To conclude, most of the current methods for identifying students at risk of drop-out make use of collected student data, including demographic, social or online learning data. The approaches based on the online learning data demonstrate the importance of capturing student learning behaviour. However, even when online learning data are not available, predictive models offer valuable insights. The success modelling can be further extended by analysing student behaviour. The approach is widely used to analyse online learning data where a sufficient amount of student-related information is available. However, we will also show that the student behaviour captured by exam success records can uncover vital information about student time management and identify patterns that lead to success.

1.2. Research goal & question

Our research focuses on first-year students of the FME, which is the most vulnerable group regarding drop-outs. First-year students face several challenges and the first exam period is one of them. In previous research [4], it has been uncovered that students' performance in the first exam period is related to their continuation in the pursuit of the academic degree. The SRL metacognitive processes reflect that successful students master their planning/time management and prioritization of their tasks [7]. We hypothesize that the order of exams in the exam period influences student success. Keeping this in mind, we formulate the research question as follows:

Research Question:*Can the student exam behaviour represented by the order of student's exams; be used as a predictor of student success in the first academic year?*

For that purpose, we formulated the task of predicting the first-year academic results using multiple predictive models. The student result acquires two possible values: *Pass* and *Fail*, where the former represents the fact that the student passed all requirements for continuing studies in the next academic year, and the latter means that the student drops out from the study. The presented approach combines the student success predictions with the analysis of student behaviour in VLEs. Since the information available is sparse, a new method for encoding student exam states, which reflects and uses it for the prediction, has been developed.

FME is a typical Czech university focusing on face-to-face education with diverse information systems and data sources containing information about the students. Thus, it is usually challenging to collect the student data in a form suitable for the analysis. Given such a heterogeneous environment, only exam performance data is available and serve as an input for the training of each predictive model. Multiple training predictive models have been employed, namely: Support Vector Machines, k-Nearest Neighbours, Classification and Regression tree, Random Forest and Naive Bayes algorithms.

Three experiments to explore whether the student exam behaviour improves the prediction of student success in the first academic year will be conducted. The first and second experiment employs the 10-fold and leave-one-out cross-validation over the dataset produced from all available data. These two experiments evaluate the predictive modelling "power" and compare the newly developed approach using the temporal information about the student exam behaviour to the baseline approach using the exam outcomes only. Finally, the third experiment will simulate a real-world scenario using the data collected in the first academic year for model training using 10-fold cross-validation and the data from the second academic year for validation. The

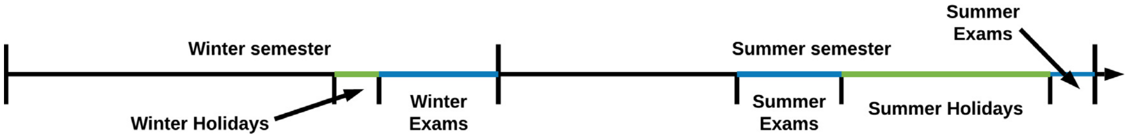

Fig. 1. Academic year at FME.

Table 1
First (winter) semester courses.

Course	Course
Mathematics I. *	Engineering Design I.
Constructive geometry *	Fundamentals of Technology I.
Physics I. *	Computer Support for Study
History of Technology	Management skills

results will show that the sequential approach enhances the models' predictive power and enables us to explore the most critical behaviour (exam-taking) patterns.

1.3. Educational setting

The Czech Technical University in Prague is one of the biggest and oldest technical universities in Europe. One of its faculties is the Faculty of Mechanical Engineering, which offers education in three bachelor and 13 master study programmes. It has approximately 3,000 students currently pursuing their degrees. About one-fourth of them are first-year students of bachelor programmes. Around 400 students register for study in the Theoretical fundamentals of the Mechanical Engineering programme, which is the faculty's flagship bachelor programme.

The study year at FME (see Fig. 1) is divided into winter and summer semesters. Each semester consists of a period with lectures and seminars and an exam period in the end. Besides, the study year includes summer holidays, typically without any educational activity. The Winter semester usually starts at the beginning of October. It has 14 study weeks with winter holidays in the last quarter. The summer semester begins after the winter exam period, and it is followed by the exam period, which ends the academic year.

To simplify the transition from high school, FME defines courses that first-year FME students are required to attend in the winter semester (Table 1). The exam period has six weeks (five regular and one extra for first-year students). Every week each course usually has one or more exam dates listed. Every student has two attempts to pass the exam, and if failing both, the dean may allow the third one. Three of the prescribed courses end with the exam. They are: Mathematics I., Constructive geometry and Physics I. The other courses can be passed by the successful completion of the laboratory tasks and seminar work. Passing at least one exam course and all non-exam courses is required to progress from the first semester. If students pass approximately half of the first years' exams, they can continue their studies. It is worth noting that the number of exams students need to pass in the following years is increasing. Thus, success in the first semester exams is a good indicator of student progression into the second academic year.

After evaluating the educational context at FME, we decided to focus on courses that end with the exam. The exam courses represent the important content of the studied area. In the following study years, these exam courses are the most relevant to the study success. Thus this paper focuses on the three courses in the winter semester, denoted by an asterisk in the table.

Table 2
General dataset statistics.

Value	2017/2018 dataset	2018/2019
Number of students	360	330
Number of failing students	140 (39%)	121 (37%)
Students who did not passed all exams	178	207

Table 3
Examples of attempt log.

Student id	Date	Course	Result
15456106	2018-01-22	Mathematics I.	F
15456106	2018-02-16	Mathematics I.	E
15456106	2018-02-24	Constructive geometry	D
15456106	2018-02-29	Physics I.	A
13274506	2018-01-26	Mathematics I.	B
...	...	...	...

2. Data

Data from two consecutive academic years, 2017/2018 and 2018/2019, has been collected. Both data sets contain information about the freshman students (new students without any previous academic history) who registered for the studies in the study programme Theoretical Fundamentals of Mechanical Engineering, representing the most prominent and major study programme at the FME. The general statistics for both data sets are available in Table 2. Note that the percentage of students who did not continue their studies in the following year is almost the same. However, in the year 2018/2019, the number of students who did not successfully finish all exams is higher than in previous years.

The exam data collected by the university are stored as log entries containing the date and the outcome of the exam attempt for every student. The randomly generated example of an attempt log is shown in Table 3. The log contains the id of the student, date of exam, course and the exam outcome. The first student (we will call him Bob) in our example took four exams in total. One can observe that Bob failed in the first attempt of the Mathematics I. exam (mark F). The following sections will further use Bobs' example for the explanation of the approach for transforming data. As a target variable, we collected information about the successful student progression to the next academic year. Students who passed the academic year and started the second study year were labelled as *Pass* and students who drop-out from the studies were marked as *Fail*.

3. Methods

This section introduces our exam progression encoding method for transforming all available contextual information into a form suitable for predictive modelling and interpretation. The experiments will then be presented, introducing the methods for model training, experiment design, and evaluation metrics. Finally, the technique for the detection of critical behaviour patterns within the sequential data will be presented.

Table 4
Examples of weekly student-exam states.

state	Physics I.	Constructive geometry	Mathematics I.
0	The exam is not attempted.	The exam is not attempted.	The exam is not attempted.
3	The exam is not attempted.	The exam is not attempted.	Passed
15	The exam is not attempted.	Passed	Passed
16	Failed 1st attempt	The exam is not attempted.	The exam is not attempted.
48	Passed	The exam is not attempted.	The exam is not attempted.
49	Passed	The exam is not attempted.	Failed 1st attempt
61	Passed	Passed	Failed 1st attempt
62	Passed	Passed	Failed 2nd attempt
63	Passed	Passed	Passed

3.1. Encoding the student exam progression

Every student needs to pass E exams at the end of the semester. For each exam students can be in a different state s_i, which represents whether the they attempted, failed or passed the exam. Thus each student might be in one of N states from set $S = \{s_1, s_2, \dots, s_N\}$, where $s_i \in \mathbb{N}$ and $s_i < s_j$, where i and j are indexes and $i < j$. Having the set of possible student states S, we can then represent student exam period state as the vector containing E elements:

$$\vec{e} = (e_1, e_2, \dots, e_E), \tag{1}$$

where each element e_i represent the student state in one exam and can acquire values from the set S. The vector contains a finite number of combinations of its elements: N^E and reflects the student's state for every exam he/she needs to pass during the exam period. Vector $\vec{e}$ can be measured at different times of the exam period.

To further reduce the dimension without loss of information the vector $\vec{e}$ can be transformed to the decimal number X:

$$X = \sum_{i=1,\dots,E} N^{i-1} e_i. \tag{2}$$

The final transformation provides us with the decimal number representing the cumulative exam period state of one student. The state accumulates all "achieved" results to the time when the sample has been taken.

In our case, FME students need to pass the three required exams, where for each of the exam, one of the following four possible states (results) can be achieved:

1. The student did not attempt the exam so far.
2. The student failed the first attempt.
3. The student failed the second attempt (and failed the course).
4. The student passed the exam (and passed the course).

Thus in our case $S = \{0, 1, 2, 3\}$, $N = 4$ and $E = 3$. This gives us $4^3 = 64$ possible student exam states for every moment of the winter exam period. The state vector containing 3 elements (one for each exam) is then:

$$\vec{e} = (e_1 = P, e_2 = C, e_3 = M), \tag{3}$$

where $P \in S$, $C \in S$ and $M \in S$ represents students' current achievement in Physics I. (P), Constructive geometry (C) and Mathematics I. (M). The vector $\vec{e}$ represents a quaternary (base-4) number, which can be transformed into decimal number X:

$$X = 16P + 4C + M. \tag{4}$$

The number $X \in \{0, 1, 2, \dots, 63\}$ then represents the student exam state, which accumulates the students' achievement in the exam period so far. $X = 0$ represents the fact that the student did not attempt any exam and $X = 63$ represents the fact that the student passed all three exams. The table (Table 4) shows the examples of possible student exam states.

3.2. Student exam sequence

As shown in the previous section, the student exam state represents student success until the selected time. During the exam period, the exam state collects each student's results and reflects their progression in time. Each student has a unique sequence of student exam states s.

For the FME, the exam period length is six weeks, and every week each exam has at least one exam date opened. That means that every student has the opportunity to attend any exam. The most usual way of sitting exams is also one per week. Thus, we decided to sample the exam period into six time periods and took the samples for each student at the end of each period. Therefore in our case, the student exam sequences will be of length six.

To illustrate this, we can use Bob as an example. His sequence using the encoding from above will be $s_{Bob} = \langle 1, 1, 1, 3, 15, 63 \rangle$. The sequence reflects that he failed the Mathematics I. exam in the first week of the exam period and then spent two weeks preparing for the other attempt. After the Mathematics I. exam, Bob successfully passed other exams in the following weeks.

The complete sequence fully characterizes the student exam behaviour. However, parts of the sequence may be shared with groups of students and may be used for recognizing the critical patterns in exam behaviour. Therefore the subsequences of different length might be extracted from the original full-length sequence and used for the predictive modelling. These subsequences are called k-grams [28], where k represents the number of consecutive symbols in sequence. The k-gram information extracted from the sequence is then used as the input for the predictive model.

3.3. Predictive modelling of student outcomes

To examine the importance of exam behaviour, we first processed the raw data and formed two kinds of data:

- Baseline dataset – contains the information about the student exam success for each course without any indication of the sequential order of exams taken. Each exam best result is numerically encoded with A corresponding to 1, B to 1.5, etc. (see Table 5 for example of the student Bob).
- Sequence datasets – each of the 6 datasets contain all possible k-grams formed from student exam sequences for one value of k. In our case of $k \in \{1, \dots, 6\}$ we formed six datasets containing $(N^E)^k$ features at maximum for each student. We do not include k-gram features with no occurrence in the original raw data. The number of features in each dataset can be seen in Table 6.

Both kinds of data, baseline and sequential, then served as the inputs in three experiments with four machine learning methods used for the predictive model training. These are k-Nearest Neighbours (KNN), Classification and Regression tree (CART); Random Forest (RF); and Non-linear Support Vector Machines with Radial Basis Function kernel (SVM-RBF).

Table 5
Example of baseline data. Feature names follows the naming convention from Section 3.1.

Student id	C	M	P
15456106	2.5	3	1
...	...	...	...

Table 6
Number of features in each k-gram sequential dataset.

k	Number of features	k	Number of features
1	25	4	85
2	58	5	65
3	80	6	26

KNN [29] is a method using the training dataset directly for classification by computing the distance between training samples t_i and new instance x. The most commonly used distance is Euclidean distance, defined as:

$$d(x, t_i) = \sqrt{\sum_{j=1}^{n}(x_j - t_{ij})^2}, \tag{5}$$

where n is the number of features. Next, the k samples with the lowest distances are selected, and the class is determined using the majority voting principle. KNN method does not require any hyper-parameter tuning since it uses the data itself for classification. The only option the user needs to select is the number of neighbours k used for determining the class of the new sample. It provides a straightforward and versatile approach for classification. However, the method struggles with the so-called curse of dimensionality as the algorithm slows down with the high number of samples. Also, the method is sensitive to the outliers in the training data.

CART algorithm [30] builds the binary decision tree using Gini's impurity index, which measures how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the set:

$$G(S) = \sum_{i=1}^{m} p_i(1 - p_i), \tag{6}$$

where p_i is the probability of each class being chosen, $1-p_i$ is the probability of a mistake in class labelling, and m is the number of classes. In every iteration of the tree building procedure, the algorithm selects the feature with the lowest Gini's impurity index for splitting the instances. The procedure continues until all tree leaves contain only instances of one class. During the tree building, the complexity parameter cp is used to define the threshold for discarding node expansions not achieving minimal improvement of the criteria. After the tree is built, the post-pruning technique is deployed using cost complexity pruning criteria:

$$cc(T, t) = \frac{err(prune(T, t), S) - err(T, S)}{|leaves(T)| - |leaves(prune(T, t))|}, \tag{7}$$

where the $prune(T, t)$ is the subtree created from the tree T by "removal" of node t, S is the set of training samples, and the $|leaves()|$ is the number of leaves of the corresponding tree. The pruning continues until the complexity criteria are improving. The decision tree itself provides users with an explanation of each decision, and it is simple to understand by a wide audience. Decision tree models are sensitive to the input data, and even a small change in the data can result in a large change in the final model.

RF [31] is a tree ensemble method developed to overcome the disadvantages of individual decision tree models. It combines the predictions of multiple models to make them more accurate prediction. RF uses the variant of an ensemble algorithm called bagging. During the training, random subsamples with the replacement of the dataset are produced and used to train one model. Next, all trained models are voting for the predicted class during the prediction step, and the major vote is taken. The voting reduces the overfitting of each model, and with a large number of models, the accuracy increases. The RF uses instead of CART Random Tree models, which have grown "randomness" during the model training by selecting only a subset of features that can be used for the split during the tree building. The number of features m_{trn} used in the tree building procedure is the optimization parameter.

SVM-RBF is an algorithm that searches the linear classifier minimizing structural risk. The linear classifier corresponds to the decision function: $f(x) = sign(< \mathbf{w}, \mathbf{x} > +b)$, where $\mathbf{w}$ and b represent the parameters of the decision hyper-plane in the feature space, $\mathbf{x}$ represents the data vector and $< \mathbf{w}, \mathbf{x} >$ represents the inner product between hyper-plane weights and the data vector. The algorithm then maximizes the separation margin between the two classes. The margin maximization can be solved using the constrained quadratic optimization problem, whose solution is $w = \sum_i \alpha_i \mathbf{x}$. It can be shown that only samples on the decision boundary have non-zero α_i coefficients and they are called support vectors. The expansion of the algorithm called soft margin classifier, which introduces a penalty C for the sample being on the "wrong" side of the decision boundary, can be used on non-separable data. The whole problem is then formalized using L_2-norm as:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i, \\ &\text{subject to } y_i(< \mathbf{w}, \mathbf{x} > +b) \geq 1 - \xi_i \quad (i = 1, \ldots, m) \\ &\qquad\qquad \xi_i \geq 0 \quad (i = 1, \ldots, m), \end{aligned} \tag{8}$$

where m is the number of training samples, $y_i = \pm 1$, ξ_i is the penalization function and C is the cost of being on the "wrong" side of the decision boundary. Besides, many problems are not linearly separable, however, it is possible to find the dimension in which the instances are separable. To do that one can introduce the so-called kernel so that the feature space is mapped to the higher dimension, in which the samples are separable. The common kernel used is Radial Basis Function kernel: $k(\mathbf{x}, \mathbf{x}') = exp(-\sigma\|\mathbf{x} - \mathbf{x}'\|^2)$, where $\mathbf{x}$ and $\mathbf{x}'$ represent two samples and the σ is the parameter of the kernel. SVM-RBF is efficient in high dimensional spaces thanks to using kernel trick. Moreover, the algorithm is memory efficient. On the other side, it is not suitable for large data sets and underperforms in cases with higher dimensional data and fewer samples than the number of dimensions. By using the algorithm, one needs to adjust two parameters C and σ.

For answering the research question, three different experiments have been conducted. The first and second experiment uses all available data. Both experiments evaluate the predictive modelling methods using 10-fold stratified sampling cross-validation (first experiment) and leave-one-out cross-validation (second experiment). In model training, the 10-fold cross-validation (with stratified sampling) over the training data (90% of data in the first experiment, all data except one sample in the second experiment) for determining the optimal model parameters have been used. During the training phase, the following parameters have been searched: the number of neighbours k for the KNN model; the cp parameter for the CART model; m_trn for RF; and C parameter for SVM-RBF. Additionally, the SVM-RBF σ parameter has been estimated from the training samples as the

Table 7
Precision, Recall and F-measure values for class*Pass* in the first experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	F-measure	0.82 ± 0.03	0.80 ± 0.04	0.82 ± 0.02	0.82 ± 0.03	0.78 ± 0.04	0.78 ± 0.02	0.77 ± 0.01
	Precision	0.77 ± 0.04	0.76 ± 0.04	0.77 ± 0.04	0.74 ± 0.03	0.67 ± 0.03	0.65 ± 0.01	0.63 ± 0.01
	Recall	0.88 ± 0.04	0.84 ± 0.04	0.87 ± 0.04	0.92 ± 0.04	0.93 ± 0.06	0.98 ± 0.03	0.99 ± 0.01
CART	F-measure	0.82 ± 0.03	0.78 ± 0.04	0.80 ± 0.02	0.80 ± 0.02	0.79 ± 0.02	0.78 ± 0.01	0.77 ± 0.01
	Precision	0.76 ± 0.03	0.80 ± 0.04	0.74 ± 0.03	0.70 ± 0.02	0.67 ± 0.02	0.64 ± 0.01	0.63 ± 0.01
	Recall	0.89 ± 0.04	0.76 ± 0.06	0.89 ± 0.03	0.92 ± 0.03	0.95 ± 0.03	0.98 ± 0.03	1.00 ± 0.00
RF	F-measure	0.83 ± 0.03	0.81 ± 0.03	0.82 ± 0.04	0.81 ± 0.05	0.80 ± 0.05	0.78 ± 0.02	0.78 ± 0.01
	Precision	0.78 ± 0.04	0.76 ± 0.04	0.78 ± 0.04	0.77 ± 0.05	0.72 ± 0.04	0.67 ± 0.02	0.64 ± 0.02
	Recall	0.88 ± 0.03	0.88 ± 0.05	0.86 ± 0.04	0.85 ± 0.06	0.90 ± 0.07	0.94 ± 0.04	0.98 ± 0.01
SVM-RBF	F-measure	0.83 ± 0.03	0.82 ± 0.03	0.82 ± 0.02	0.80 ± 0.04	0.81 ± 0.04	0.79 ± 0.01	0.77 ± 0.01
	Precision	0.75 ± 0.03	0.78 ± 0.04	0.79 ± 0.03	0.75 ± 0.03	0.74 ± 0.04	0.68 ± 0.02	0.64 ± 0.01
	Recall	0.92 ± 0.04	0.87 ± 0.03	0.86 ± 0.05	0.86 ± 0.05	0.90 ± 0.04	0.95 ± 0.03	0.99 ± 0.01

mean value of the 0.1 and 0.9 percentile of the distribution of values calculated as L_2-norm of the differences between all samples: $\|\mathbf{x} - \mathbf{x}'\|^2$. In practice, any value lying between the percentiles provides a good option for the σ parameter [32]. Finally, the performance of the methods is measured using data from the validation set. The last experiment was performed using the data from the second academic year as the validation dataset and the data from the first academic year as the training dataset. Again, the optimal parameters of each model have been determined using 10-fold cross-validation, and the performance is measured using the validation dataset.

3.4. Evaluation measures

For the evaluation and comparison of the model performance, statistical measures Precision (*Prec*), Recall (*Rec*) and F-measure (F_{meas}) have been used [32]. All measures range from 0 to 1, where 1 represents the "full" outcome and the most desired result. $F - meas$ is computed as the harmonic mean between *Prec* and *Recall*.

In the first experiment, the cross-validation produced ten different outcomes. Thus, the mean and standard deviation of each measure will be reported. For the second experiment, the measures themselves have been reported. The leave-one-out cross-validation evaluates the predictions for each sample separately, and the results are then collected into a contingency table. In the final experiment, the advantage of a large validation dataset has been used, and the 95% confidence intervals (CI) for *Prec* and *Rec* using Wilson score interval method [33] have been estimated. The estimation of CI is possible since the *Prec* and *Rec* represent the proportions calculated from the validation dataset [34].

3.5. Important sequences detection

As the last step, we used the results of predictive modelling to identify the optimal sequence length and determine the most critical sequences regarding student exam behaviour.

For sequence identification, the Maximum Relevance and Minimum Redundancy (MRMR) algorithm [35] has been used. The algorithm is widely used within the machine learning community and balances selecting the most informative attributes and attributes with the lowest redundancy. The algorithm takes a set of attributes and a target variable as input. The next step starts with the estimation of the distribution of each attribute and the target variable. Then computes the mutual information between attributes and between attributes and the target variable. After the initial computation, the algorithm proceeds iteratively by selecting one attribute every step. Selection is based on the difference between relevancy and redundancy criteria. The positive value of criteria means that the considered attribute is more relevant to the target variable than redundant to the previously selected attributes. The negative value means the opposite – the attribute does not provide new enough information compared to the redundancy with the previously selected attributes. So that the attributes justifying the following criteria are then selected:

- attribute is the most relevant to the target variable;
- attribute is minimally redundant with all other not yet selected attributes.

The approach will choose the optimal set of attributes containing only the most relevant information regarding the outcome. After the sequential predictive models' evaluation, the k-grams of optimal length will be fed into the MRMR algorithm, and the algorithm will select the most relevant patterns.

4. Results & discussion

The following section covers the results of the presented research. The results of each experiment will be introduced and discussed. Next, the detected behavioural patterns will be presented.

We do not report the accuracy over the validation data, but this measure values varied around the value of 0.77±0.05 regardless of the selected model and used dataset. It will be shown in the following sections that by introducing the sequential data the slight reduction of the success rate for the *Pass* class will greatly increase the performance of the model over the *Fail* class.

The following sections will discuss the results in depicted in Tables 7–14. The format of the tables is the same. The first column contains the information on the method used for the predictive model training, the second column reports the measure. Next are the reported measures for baseline data followed by six columns containing the measures for models trained using the sequential type of data with the varying sequence length from 1 to 6.

4.1. Experiment 1 – 10-fold cross-validation over the entire data

Tables 7 and 8 report the results of each machine learning method evaluated using 10-fold cross-validation over all available data. The reported measures are calculated for the validation set. All reported values are in the form *mean±standard deviation*. The prediction of *Pass* students (Table 7) is more or less consistent over the all used methods with the minimal difference in all evaluation measures. We can observe that the overall F measure is slightly decreasing with the growing k-gram sequence length. The models are trained with too specific data covering only very few samples, thus they are not capable of proper generalization and tend to put all new samples to the majority class, which is in our case the *Pass* class. This trend can be especially seen

Table 8
Precision, Recall and F-measure values for class *Fail* in the first experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	F-measure	0.35 ± 0.18	0.61 ± 0.08	0.63 ± 0.07	0.57 ± 0.07	0.37 ± 0.10	0.22 ± 0.07	0.10 ± 0.04
	Precision	0.46 ± 0.19	0.68 ± 0.08	0.73 ± 0.07	0.78 ± 0.10	0.75 ± 0.17	0.80 ± 0.24	0.82 ± 0.34
	Recall	0.29 ± 0.16	0.55 ± 0.09	0.57 ± 0.10	0.46 ± 0.08	0.26 ± 0.09	0.13 ± 0.04	0.05 ± 0.03
CART	F-measure	0.32 ± 0.16	0.66 ± 0.05	0.57 ± 0.07	0.48 ± 0.06	0.34 ± 0.11	0.20 ± 0.06	0.07 ± 0.00
	Precision	0.46 ± 0.18	0.64 ± 0.06	0.72 ± 0.07	0.73 ± 0.07	0.77 ± 0.15	0.77 ± 0.25	1.00 ± 0.00
	Recall	0.25 ± 0.14	0.69 ± 0.07	0.48 ± 0.08	0.36 ± 0.06	0.23 ± 0.09	0.10 ± 0.05	0.02 ± 0.02
RF	F-measure	0.40 ± 0.16	0.61 ± 0.08	0.66 ± 0.08	0.63 ± 0.10	0.54 ± 0.09	0.36 ± 0.09	0.17 ± 0.08
	Precision	0.50 ± 0.16	0.73 ± 0.08	0.73 ± 0.08	0.70 ± 0.10	0.73 ± 0.15	0.71 ± 0.17	0.70 ± 0.31
	Recall	0.34 ± 0.15	0.54 ± 0.11	0.60 ± 0.10	0.58 ± 0.11	0.43 ± 0.09	0.25 ± 0.07	0.09 ± 0.05
SVM-RBF	F-measure	0.30 ± 0.13	0.65 ± 0.07	0.66 ± 0.05	0.60 ± 0.07	0.57 ± 0.08	0.40 ± 0.07	0.14 ± 0.07
	Precision	0.52 ± 0.21	0.74 ± 0.06	0.74 ± 0.06	0.70 ± 0.08	0.75 ± 0.10	0.77 ± 0.11	0.76 ± 0.33
	Recall	0.22 ± 0.10	0.59 ± 0.09	0.61 ± 0.07	0.52 ± 0.07	0.46 ± 0.09	0.28 ± 0.07	0.07 ± 0.05

with the CART model sensitive to changes in the input data. The same behaviour can be observed for the students class *Fail*. However, there is a difference when using the baseline dataset and sequential datasets, especially with lower *k*-grams. This difference is approximately 0.3 improvement when using sequence data of length 1 or 2 across all models. The improvement might be interpreted as a strong argument for a proper exam taking plan when suggesting time management strategies to students. The similar performance declining trend with increased length of sequence as in the case of *Pass* students can be observed also within the *Fail* students. The decline again points towards the specific description of the student in sense of features, which leads to the underfitting of the model. The CART model is the most sensitive and in the case of a sequence of length 6 it nearly detects none of the *Fail* cases. The standard deviation of reported measures reduced significantly when using the sequential data. This reduction is an added value of the use of sequential data as the stability of results makes the model less vulnerable to the change of training dataset. The predictive models trained over the sequence of length 2 data reported the best F-measure almost in every model for both classes. The only exception is the CART model, which is underperforming in comparison to other models suggesting that the model suffers from its sensitivity to the training data.

4.2. Experiment 2 – leave-one-out cross-validation over the entire data

The results of the experiment are presented in Tables 9 and 10. The reported results are computed over the contingency table from all 690 cross-validation steps since each step produce the prediction on only one unseen sample. Again, the reported F-measure for the *Pass* student class is more or less similar across all the datasets and used models with small variation for the CART model and data for the sequence of length 1. The reported performance decline with the length of sequences is presented also in this experiment. For the students of class *Fail* the measures improve when using the sequential data in comparison to baseline data the improvement is being highest with sequences of lower lengths and diminishes with the longer sequences, which suggests that the data are too specific for models to optimally decide the correct class. This trend is most evident in CART and KNN models where the evident decline starts with the length of 4 in comparison to models RF and SVM-RBF where it starts with a length of 5 suggesting that the latter models are less sensitive to the features with high underlying complexity. An interesting observation is that for the CART model it is almost impossible to correctly identify the *Fail* students using a baseline dataset. This shows its high-variance for to the change of the training dataset. The "accuracy" of sequential models with lower sequence length shows the importance of detecting the desired pattern in student exam taking. Thanks to the fact that the *k*-grams contain shorter sequences it is possible to use such features at the start of the exam period to identify students with possible misconceptions in their exam taking strategies and help them with further planning.

Table 9
Precision, Recall and F-measure values for class *Pass* in the second experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	F-measure	0.82	0.81	0.82	0.81	0.80	0.78	0.77
	Precision	0.76	0.77	0.78	0.73	0.69	0.65	0.63
	Recall	0.88	0.85	0.88	0.91	0.96	0.98	0.99
CART	F-measure	0.82	0.77	0.80	0.80	0.78	0.78	0.77
	Precision	0.72	0.81	0.74	0.70	0.67	0.65	0.63
	Recall	0.96	0.73	0.89	0.92	0.95	0.98	1.00
RF	F-measure	0.82	0.82	0.82	0.81	0.81	0.79	0.77
	Precision	0.77	0.75	0.78	0.77	0.73	0.68	0.64
	Recall	0.88	0.89	0.87	0.85	0.91	0.94	0.98
SVM-RBF	F-measure	0.83	0.82	0.82	0.81	0.81	0.80	0.77
	Precision	0.76	0.78	0.78	0.75	0.74	0.69	0.64
	Recall	0.93	0.87	0.86	0.87	0.90	0.94	0.99

Table 10
Precision, Recall and F-measure values for class *Fail* in the second experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	F-measure	0.36	0.64	0.65	0.57	0.41	0.21	0.09
	Precision	0.48	0.71	0.74	0.75	0.79	0.76	0.81
	Recall	0.29	0.58	0.58	0.46	0.28	0.12	0.05
CART	F-measure	0.08	0.66	0.57	0.49	0.33	0.21	0.04
	Precision	0.29	0.62	0.72	0.73	0.74	0.78	0.86
	Recall	0.04	0.71	0.48	0.37	0.21	0.12	0.02
RF	F-measure	0.39	0.61	0.67	0.64	0.56	0.38	0.14
	Precision	0.50	0.75	0.74	0.70	0.75	0.72	0.70
	Recall	0.31	0.52	0.61	0.58	0.44	0.26	0.08
SVM-RBF	F-measure	0.31	0.66	0.66	0.61	0.58	0.44	0.14
	Precision	0.54	0.73	0.72	0.71	0.75	0.76	0.77
	Recall	0.22	0.59	0.61	0.54	0.48	0.31	0.08

4.3. Experiment 3 - second year data as validation dataset

The final experiment simulates a real-world scenario when the data availability and consistency between more than two consecutive years might be compromised due to small changes in the educational setting [36]. Thus, this experiment was designed to evaluate predictive modelling in the case when only one year of data has been used for model training. The results of the experiments are depicted in Tables 11, 12 and the corresponding confidence intervals are in Tables 7 and 14. The prediction

Table 11
Precision, Recall and F-measure values for *Pass* the third experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	F-measure	0.75	0.76	0.75	0.82	0.82	0.80	0.78
	Precision	0.80	0.76	0.78	0.77	0.73	0.69	0.63
	Recall	0.71	0.77	0.72	0.88	0.92	0.95	1.00
CART	F-measure	0.83	0.73	0.57	0.55	0.49	0.79	0.78
	Precision	0.79	0.74	0.83	0.85	0.85	0.67	0.63
	Recall	0.88	0.72	0.43	0.41	0.34	0.96	1.00
RF	F-measure	0.75	0.73	0.81	0.79	0.78	0.76	0.77
	Precision	0.80	0.77	0.74	0.80	0.80	0.73	0.65
	Recall	0.71	0.69	0.90	0.78	0.77	0.80	0.95
SVM-RBF	F-measure	0.75	0.80	0.81	0.53	0.43	0.67	0.77
	Precision	0.80	0.78	0.78	0.86	0.85	0.72	0.65
	Recall	0.71	0.82	0.84	0.39	0.29	0.62	0.95

Table 12
Precision, Recall and F-measure values for *Fail* the third experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	F-measure	0.39	0.59	0.61	0.61	0.53	0.39	NA
	Precision	0.34	0.59	0.57	0.71	0.76	0.74	NA
	Recall	0.46	0.59	0.64	0.54	0.41	0.26	0.00
CART	F-measure	0.36	0.54	0.60	0.61	0.59	0.30	NA
	Precision	0.45	0.53	0.46	0.46	0.44	0.72	NA
	Recall	0.31	0.55	0.85	0.88	0.89	0.19	0.00
RF	F-measure	0.39	0.59	0.56	0.66	0.64	0.54	0.19
	Precision	0.34	0.55	0.73	0.64	0.62	0.59	0.58
	Recall	0.46	0.64	0.45	0.67	0.66	0.50	0.12
SVM-RBF	F-measure	0.39	0.63	0.62	0.61	0.58	0.53	0.19
	Precision	0.34	0.66	0.68	0.46	0.43	0.47	0.58
	Recall	0.46	0.60	0.58	0.89	0.91	0.60	0.12

of *Pass* students shows similar trends as in the previous two experiments. It shows relatively stable results over the datasets regardless of the used modelling method. In this experiment, the CART and SVM-RBF models showed a decline of the F-measure for sequences of length 3 and 4, which was caused by capturing significantly fewer samples of the *Pass* class resulting in a lowered recall. The "fluctuation" is due to the lowered proportion between feature dimension and the diversity of the data reflected by the number of training samples leading to improper fitting of the models. The "fluctuation" suggests that the data between years are not the same in the sense of the educational context, and further investigation is needed to uncover the reasons. Next, the prediction performance for the class *Fail* shows similar trends as in previous examples. The improvement in F-measure is significant between the model based on the baseline data and the models using sequential data. Especially for the sequences of shorter lengths. The declining trend of performance for the class *Fail* is mostly evident for the KNN and CART classifiers when for the sequence of length six, the models were unable to identify any *Fail* students. Again this decline is the manifestation of specific data, which causes the improper fitting of the underlying models.

4.4. Summary of experiments

The three experiments highlighted several findings, which manifests via the comparison of predictive modelling outcomes with a different set of features used to train the predictive model (baseline data or temporal sequence exam-taking data). The following summarizes the main findings:

- **The sequential models outperform the baseline model for the group of *Fail* students**. All three experiments support the evidence that sequential data (exam-taking sequences) improves the success prediction outcome. The improvement is for the *Fail* students class, while for the *Pass* student class is the outcome almost unchanged. The observation implies that the predictive models based on the exam-taking patterns provide better insight into at-risk student behaviour while keeping the "quite right" identification almost unchanged. The identification of at-risk students in the *Fail* class is important because it provides the teaching staff with the opportunity to carry out the intervention way before the student drops out of the studies. And thanks to the fact that is identifying the "quite right" students is unchanged, the intervention efforts can focus fully on students in need of additional support.
- **The models based on shorter exam sequences improve the prediction more than models based on long exam sequences**. The shorter sequences capture the shorter periods and represent "smaller" fluctuations in the behaviour of the student exam-taking patterns. Thus short sequences provide more "general patterns" of exam behaviour than longer sequences, specific to the individual students, mainly because the long exam-taking patterns represent the whole exam period. Thus, using short sequence exam behaviour patterns leads to improved prediction, while the longer sequence is not because of its specificity. Moreover, the student data sample is not large enough to capture all important combinations of behavioural patterns. In addition, the short sequences also enable the intervention sooner because the time required to measure the sequence patterns is lower.
- **There was no difference between used machine learning algorithms for model building**. Almost all results show nearly the same outcome measured by the statistical measures regardless of the used algorithm. The insignificance of the used method is caused by the fact that the student cohort is relatively small compared to the captured information. Thus the used data is more important than the used predictive model. More advanced models such as neural network will also be limited due to the same fact of small number of measured instances. This show the importance of the careful selection of used monitoring attributes for a similar type of task, where the sparsely measured data represent the small number of cases.
- **For the real-world case (third experiment), the models trained using sequential exam behavioural data perform better than the baseline model**. The baseline dataset (using only exam results) does not reflect that the student passes the exam on the second attempt and the order in which the student took the exams. Thus, the baseline model does not use all available information in the data, resulting in significantly lower performance when only a few samples are used for the model training. Using data from multiple consecutive years can reduce the issue. However, the changes in the teaching during the time can introduce a different kind of error. In comparison to the baseline models, the sequential models result in relatively stable results on unseen data.

4.5. Detection of important sequences

Based on the evaluation, the exam behaviour sequence of length 2 is the most descriptive. Thus, the MRMR algorithm has been applied to the corresponding data to detect the most essential student exam behaviour patterns. From 4096 (63^2) possible sequences, only 56 are presented in a real-world dataset. The presence of only such a small number of patterns can be interpreted by sharing the typical behavioural patterns between students in explaining, especially on a short-term scale. Out of

Table 13
Precision, Recall confidence intervals for *Pass* the third experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	precision	[0.75, 0.85]	[0.71, 0.81]	[0.73, 0.83]	[0.72, 0.81]	[0.69, 0.78]	[0.64, 0.74]	[0.58, 0.68]
	recall	[0.66, 0.76]	[0.72, 0.81]	[0.67, 0.77]	[0.83, 0.92]	[0.88, 0.97]	[0.9, 0.99]	[0.95, 1.05]
CART	precision	[0.75, 0.84]	[0.69, 0.78]	[0.78, 0.88]	[0.80, 0.90]	[0.80, 0.90]	[0.62, 0.72]	[0.58, 0.68]
	recall	[0.83, 0.92]	[0.67, 0.77]	[0.38, 0.48]	[0.36, 0.46]	[0.29, 0.40]	[0.91, 1.01]	[0.95, 1.05]
RF	precision	[0.75, 0.85]	[0.72, 0.82]	[0.70, 0.79]	[0.76, 0.85]	[0.75, 0.84]	[0.68, 0.78]	[0.60, 0.70]
	recall	[0.66, 0.76]	[0.65, 0.74]	[0.86, 0.95]	[0.74, 0.83]	[0.72, 0.82]	[0.75, 0.85]	[0.90, 1.00]
SVM-RBF	precision	[0.75, 0.85]	[0.74, 0.83]	[0.73, 0.82]	[0.81, 0.91]	[0.80, 0.90]	[0.67, 0.78]	[0.60, 0.70]
	recall	[0.66, 0.76]	[0.78, 0.87]	[0.80, 0.89]	[0.34, 0.44]	[0.24, 0.34]	[0.57, 0.67]	[0.90, 1.00]

Table 14
Precision, Recall confidence intervals for *Fail* the third experiment.

Method	Measure	Baseline	Sequence length					
			1	2	3	4	5	6
KNN	precision	[0.29, 0.39]	[0.54, 0.64]	[0.53, 0.62]	[0.67, 0.76]	[0.71, 0.8]	[0.7, 0.79]	NA
	recall	[0.41, 0.51]	[0.54, 0.63]	[0.60, 0.69]	[0.49, 0.58]	[0.37, 0.46]	[0.22, 0.31]	[0.00, 0.05]
CART	precision	[0.40, 0.50]	[0.48, 0.58]	[0.41, 0.51]	[0.41, 0.51]	[0.39, 0.49]	[0.67, 0.77]	NA
	recall	[0.26, 0.35]	[0.50, 0.60]	[0.80, 0.90]	[0.83, 0.93]	[0.84, 0.94]	[0.14, 0.24]	[0.00, 0.05]
RF	precision	[0.29, 0.39]	[0.5, 0.59]	[0.69, 0.78]	[0.60, 0.69]	[0.58, 0.67]	[0.54, 0.64]	[0.53, 0.63]
	recall	[0.41, 0.51]	[0.59, 0.68]	[0.41, 0.50]	[0.62, 0.71]	[0.62, 0.71]	[0.45, 0.54]	[0.07, 0.17]
SVM-RBF	precision	[0.29, 0.39]	[0.62, 0.71]	[0.63, 0.72]	[0.41, 0.51]	[0.37, 0.48]	[0.42, 0.52]	[0.53, 0.63]
	recall	[0.41, 0.51]	[0.56, 0.65]	[0.53, 0.62]	[0.84, 0.94]	[0.86, 0.96]	[0.54, 0.65]	[0.07, 0.17]

56, only 38 *k*-grams have the values of MRMR criterion positive (more relevant than redundant). The negative criterion value reflects that the *k*-gram does not provide enough additional information and creates redundancy in the data without any added value. Thus we removed the negative criteria attributes from the further analysis. From the 38 *k*-grams we selected *n* the first *k*-grams with the MRMR criterion value greater than 10% of the most significant MRMR criterion value (the value of the first selected most influential *k*-gram). The selection process results in the analysis of the top eight sequence combinations (see Table 15) with the most relevant information regarding the student outcome:

- **63–63** - the most obvious combination, which represents those students who finished all the exams and have nothing to do. However, an interesting observation is that students need to finish the exam in week five latest. We also observed the fact that no student has finished the exams earlier than in three weeks. Both facts suggest that successful students follow only a few exam strategies. This combination is the strongest predictor of future academic success.
- **15–15** - combination is the "predecessor" of the previous one. It represented the state when the student passed Mathematics I. and Constructive geometry. It also reflects that students leave the Physics I. exam to be the last, focusing on less challenging course exams.
- **0–4** - combination is the strongest predictor of failure. It represents the states when the student did "nothing" in one week and failed the Constructive geometry exam in the following week.
- **0–0** - another failure predictor. Students who have this pattern did not attempt any exam in two consecutive weeks. Keeping in mind that the exam period has six weeks, one can expect problems in passing the exams under time pressure.
- **1–17** - pattern represents students who failed the first attempt of Mathematics I. in one week and Physics I. in the following week. Since Physics I. is the most challenging exam and the "requirement" is the basic understanding of the underlying mathematics. It can be suggested that students with this pattern in their exam behaviour sequence are more prone to failure and need improvement of their time management skills.
- **4–5** - the pattern is similar to the previous one. It represents a failure in Constructive geometry in one week followed by a loss in Mathematics I. Again, this shows the bad habit of switching the exam topics often without proper preparation.
- **3–15** - is the predictor of success and represents success in Mathematics I. followed by success in Constructive geometry. This sequence is typical in the early weeks of the exam period for successful students.
- **21–21** - sequence representing exam "paralysis". The student failed their first attempts in all three exams, and now he does nothing between weeks. The pause, in this case, is not desirable since these states appear in the late weeks of the exam period, and there is not much time for improvement.

5. Conclusions

In this paper, we focused on answering the question: *Can the student exam behaviour represented by the order of student's exams; be used as a predictor of student success in the first academic year?*, in the context of classical brick-and-mortar Czech university. The exam taking patterns manifest one aspect of the metacognitive SRL process, especially the planning and time management of the student. Exam taking is recognized by teachers as a critical skill, which every student of the FME needs to master. Thus the formulated task aimed at first academic year student success prediction based on the exam taking behaviour. During the first academic year the most educational settings are fixed except the exam periods, which can students, within some constraints, plan themselves. The student exam progression has been encoded, and for each student, the exam sequence has been calculated. To evaluate the added value of sequential data and explore typical patterns of student exam behaviour four types of machine learning methods have been used for training the models within three particular experiments. In addition, the most successful sequence length has been used for the analysis of typical exam-taking

Table 15
MRMR score values for each *k*-gram. Bold text indicates the selected combinations for the analysis.

State 1	State 2	MRMR	State 1	State 2	MRMR
63	**63**	**0.107**	12	60	0.003
15	**15**	**0.019**	0	16	0.003
0	**4**	**0.016**	1	5	0.003
0	**0**	**0.020**	1	21	0.003
1	**17**	**0.014**	16	19	0.003
4	**5**	**0.013**	22	25	0.002
3	**15**	**0.015**	3	3	0.002
21	**21**	**0.012**	25	25	0.001
16	16	0.010	5	21	0.001
1	1	0.011	16	2	0.001
3	7	0.010	60	60	0.001
15	63	0.010	25	63	0.002
17	21	0.009	16	20	0.001
3	25	0.009	4	4	0.001
0	13	0.006	60	63	0.001
0	1	0.006	3	23	0.001
0	17	0.005	7	7	0.001
19	23	0.004	0	5	0.000
20	21	0.003	0	20	0.000

behavioural patterns. Based on the results of all three experiments, it can be concluded that the sequential data-based models with the sequence of length 2 provide the best performance, and the improvement is significant compared to the baseline predictions using only exam results. It has also been observed that the cumulative states themselves offer more information and increased in performance of predicting the at-risk (failing) students compared to the baseline model. The increase makes sense since the cumulative state's information is richer than the information in the exam result only. In contrast to the failing student detection, detection of passing students is more accurate for both baseline and sequential predictive models because successful students tend to use fewer exam strategies, which is reflected in the diversity of sequences. It has also been shown that there are basic exam behaviour patterns within the data, which are desirable to detect and can serve as triggers for targeted help. The most important patterns are of length 2, which represents student exam planning behaviour in two consecutive weeks. If students behave by following the "wrong" patterns, they have a much higher probability of failing their study. The research focused on weekly aggregated behaviour, which means that the necessary intervention might be carried out lately. However, the detection is performed in the winter semester. The faculty staff might have the whole summer semester to intervene with the students who fulfil the criteria for continuing the studies in the summer semester. In this research, we did not focus on the *k*-gram exact position in the student exam sequence, and we focused only on detecting the *k*-gram presence. That means that some patterns of behaviour might appear in the data later in the exam period. However, essential predictors of failure (such as doing nothing in two consecutive weeks) are presented in the data from the beginning of the exam period. Besides the reported results, FME is conducting interventions on a group of low performing students based on the reported findings by conducting seminars and lectures in which they provide the freshman students with the essential aspects regarding the exam period organization and planning. In addition to that, FME plans to conduct student mentoring on the group of the most at-risk students. We plan to explore the causal aspects further manifested as the different exam-taking patterns shown in this paper by conducting qualitative pedagogical research within first-year students.

CRediT authorship contribution statement

Jakub Kuzilek: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Zdenek Zdrahal:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Viktor Fuglik:** Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by a junior grant research project by Czech Science Foundation GACR no. GJ18-04150Y.

References

[1] The log-on degree; technology and universities, The Economist (US) (2015) URL https://www.economist.com/united-states/2015/03/12/the-log-on-degree.

[2] J. Quinn, Drop-out and completion in higher education in Europe, 2013, European Union.

[3] R. Kabra, R. Bichkar, Performance prediction of engineering students using decision trees , Int. J. Comput. Appl. (0975 - 8887) 36 (11) (2011).

[4] Z. Zdrahal, M. Hlosta, J. Kuzilek, Analysing performace of first year engineering students, in: Proceedings of the Sixth International Conference on Learning Analytics and Knowledge, LAK '16, 2016.

[5] E. Seymour, N.M. Hewitt, Talking About Leaving, Westview Press, Boulder, CO, 1997.

[6] P.A. Daempfle, An analysis of the high attrition rates among first year college science, math, and engineering majors, J. College Stud. Retent.: Res. Theory Pract. 5 (1) (2003) 37–52.

[7] A.J. Sebesta, E. Bray Speth, How should I study for the exam? Self-regulated learning strategies and achievement in introductory biology, CBE—Life Sci. Educ. 16 (2) (2017) ar30.

[8] B.J. Zimmerman, Models of self-regulated learning and academic achievement, in: Self-Regulated Learning and Academic Achievement, Springer, 1989, pp. 1–25.

[9] Z. Papamitsiou, A.A. Economides, Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence, Educ. Technol. Soc. 17 (4) (2014) 49–64.

[10] P. Van der Plank, et al., Effects of Habsburg educational policies measured by census statistics, Jezikoslovlje 13 (2) (2012) 373–393.

[11] Z. Zdrahal, M. Hlosta, J. Kuzilek, Analysing performance of first year engineering students, in: Learning Analytics and Knowledge: Data Literacy for Learning Analytics Workshop, 2016, URL http://oro.open.ac.uk/58597/.

[12] K.E. Arnold, M.D. Pistilli, Course signals at purdue: Using learning analytics to increase student success, in: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 267–270, http://dx.doi.org/10.1145/2330601.2330666.

[13] G. Kennedy, C. Coffrin, P. de Barba, L. Corrin, Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance, in: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, LAK '15, 2015, pp. 136–140, http://dx.doi.org/10.1145/2723576.2723593, URL http://dl.acm.org/citation.cfm?id=2723576.2723593.

[14] B. Baradwaj, S. Pal, Mining educational data to analyze student's performance, Int. J. Od Adv. Comput. Sci. Appl. 2 (6) (2012) 63–69, arXiv:1201.3417, URL http://arxiv.org/abs/1201.3417.

[15] S. Shehata, K.E. Arnold, Measuring student success using predictive engine, in: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, LAK '15, 2015, pp. 416–417, http://dx.doi.org/10.1145/2723576.2723661, URL http://dl.acm.org/citation.cfm?id=2723576.2723661.

[16] A. Wolff, Z. Zdrahal, A. Nikolov, M. Pantucek, Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment, in: LAK '13 Proceedings of the Third International Conference on Learning Analytics and Knowledge, 2013, pp. 145–149.

[17] A. Wolff, Z. Zdrahal, A. Nikolov, M. Pantucek, Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment, in: Proceedings of the Third International Conference on Learning Analytics and Knowledge, LAK '13, 2013, pp. 145–149, http://dx.doi.org/10.1145/2460296.2460324.

[18] A. Wolff, Z. Zdrahal, D. Herrmannova, J. Kuzilek, M. Hlosta, Developing predictive models for early detection of at-risk students on distance learning modules, in: 4th International Conference on Learning Analytics and Knowledge, Vol. 1137, LAK 2014, 2014, p. 4, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84924940432&partnerID=tZOtx3y1.

[19] M. Xenos, Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks, Comput. Educ. 43 (4) (2004) 345–359, http://dx.doi.org/10.1016/j.compedu.2003.09.005.

[20] E. Howard, M. Meehan, A. Parnell, Contrasting prediction methods for early warning systems at undergraduate level, Internet Higher Educ. 37 (2018) 66–75, http://dx.doi.org/10.1016/j.iheduc.2018.02.001, URL http://www.sciencedirect.com/science/article/pii/S1096751617303974.

[21] S. Huang, N. Fang, Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, Comput. Educ. 61 (2013) 133–145, http://dx.doi.org/10.1016/j.compedu.2012.08.015.

[22] A. Sharabiani, F. Karim, A. Sharabiani, M. Atanasov, H. Darabi, An enhanced bayesian network model for prediction of students' academic performance in engineering programs, in: 2014 IEEE Global Engineering Education Conference, EDUCON, 2014, pp. 832–837, http://dx.doi.org/10.1109/EDUCON.2014.6826192, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6826192.

[23] M. Pandey, V.K. Sharma, A decision tree algorithm pertaining to the student performance analysis and prediction, Int. J. Comput. Appl. 61 (13) (2013).

[24] M. Hlosta, D. Herrmannova, L. Vachova, J. Kuzilek, Z. Zdrahal, A. Wolff, Modelling student online behaviour in a virtual learning environment, LAK (2014) 2–5, URL http://ceur-ws.org/Vol-1137/LA_machinelearning_submission_4.pdf.

[25] F. Okubo, A. Shimada, Y. Taniguchi, A visualization system for predicting learning activities using state transition graphs, in: 14th International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2017, 2017, pp. 173–180.

[26] D. Davis, G. Chen, C. Hauff, G.-J. Houben, Gauging MOOC learners' adherence to the designed learning path, Int. Educ. Data Min. Soc. (2016).

[27] C. Kent, C.A. Boulton, H. Williams, Towards measurement of the relationship between student engagement and learning outcomes at a Bricks-and-Mortar university, in: Joint Proceedings of the Sixth Multimodal Learning Analytics (MMLA) Workshop and the Second Cross-LAK Workshop, MMLA-CrossLAK, 2017, pp. 4–14, URL https://ore.exeter.ac.uk/repository/bitstream/handle/10871/28538/invited_paper_2.pdf.

[28] Z. Xing, J. Pei, E. Keogh, A brief survey on sequence classification, SIGKDD Explor. Newsl. 12 (1) (2010) 40–48, http://dx.doi.org/10.1145/1882471.1882478, URL https://doi.org/10.1145/1882471.1882478.

[29] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification and Scene Analysis, Vol. 3, Wiley New York, 1973.

[30] L. Breiman, J. Friedman, C. Stone, R. Olshen, Classification and Regression Trees, Taylor & Francis, 1984, URL https://books.google.de/books?id=JwQx-WOmSyQC.

[31] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324, URL https://doi.org/10.1023/A:1010933404324.

[32] M. Kuhn, K. Johnson, et al., Applied Predictive Modeling, Vol. 26, Springer, 2013.

[33] E.B. Wilson, Probable inference, the law of succession, and statistical inference, J. Amer. Statist. Assoc. 22 (158) (1927) 209–212, URL http://www.jstor.org/stable/2276774.

[34] M. Gardner, D. Altman, Calculating confidence intervals for proportions and their differences, in: Statistics with Confidence, BMJ Publishing Group, London, 1989, pp. 28–33.

[35] Hanchuan Peng, Fuhui Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[36] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, J. Vaclavek, A. Wolff, LAK15 case study 1: OU analyse: Analysing at-risk students at the open university - learning analytics review, Learn. Anal. Rev. (2015) URL http://www.laceproject.eu/learning-analytics-review/analysing-at-risk-students-at-open-university/.

Jakub Kuzilek is a senior researcher at Humboldt-Universität zu Berlin. He received his PhD in Artificial Intelligence and biocybernetics at the Faculty of Electrical Engineering of Czech Technical University in Prague in 2013. Between years 2013 and 2017, he was a research associate at Knowledge Media Institute, Open University, Milton Keynes, UK, working on the OU Analyse project, which supports OU students and lecturers via early identification of at-risk students using machine learning methods. After that, he worked at the Czech Institute of Informatics, Robotics and Cybernetics of Czech Technical University in Prague to support first-year engineering students.

Zdenek Zdrahal is Professor of Knowledge Engineering at the Open University, UK and Associate Professor at the Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University. At the Open University, he leads the OU Analyse project. His professional interests include learning analytics, machine learning and knowledge sharing.

Viktor Fuglik works since 2012 as an assistant professor at the Department of Information Technology and Education, Faculty of Education, Charles University in Prague. He gained a Ph.D. degree in the field of study Education with a focus on didactic of information and technical education. He participates in the guidance of lectures and seminars in bachelor and master studies and lifelong learning. He is professionally interested in the areas of evaluation and self-assessment in education using ICT instruments.