Article

Forecasting of Successful Completion of University Study Programs: Data Pre-processing and Optimization of LAMA BPO Algorithm

Aleksei Iurasov[1,2*], Artur Iurasov[3]
[1] Department of Business, Vilnius Business College, Kalvarijų g. 129-401, LT-08221 Vilnius, Lithuania
[2] Department of Business Technologies and Entrepreneurship, Faculty of Business Management, Vilnius Gediminas Technical University, Saulėtekio 11, Vilnius, Lithuania,
[3] Ltd "Nacionalinis švietimo centras", Žalgirio g. 92, LT-09303 Vilnius, Lithuania.
[*] Corresponding author, e-mail: *aleksei.iurasov@vilniustech.lt*

Received: 10 August 2022
Accepted: 30 August 2022
Online: 31 August 2022

Abstract. Lithuanian school graduates wishing to be admitted to state-funded places at universities undergo a competitive selection based on their final school and state exam grades. The problem of organizing competitive selection is that in Lithuania there are different types and scales of school knowledge assessments. Algorithm developed by LAMA BPO address this problem by adjusting grades into a single scale. But choice of final arithmetic values into which pupil's grades are converted is not justified theoretically. Proposed by the authors algorithm is a development of the LAMA BPO algorithm and allows to achieve a consistently higher accuracy of predicting learning results at the university. The higher accuracy of the models indicates a better capture of the central trend: a positive correlation between the level of performance in individual school disciplines and the results of university education in certain study programs.

JEL: C11, C45, C53, C61, I29.

Keywords: educational data analytics; educational data mining; learning analytics; post-secondary education.

Citation: Aleksei Iurasov, Artur Iurasov (2022) Forecasting of Successful Completion of University Study Programs: Data Pre-processing and Optimization of LAMA BPO Algorithm. – *Applied Business: Issues & Solutions* 1(2022)32–41 – ISSN 2783-6967.
https://doi.org/10.57005/ab.2022.1.5

Introduction

Lithuanian school graduates have a large variety of study programs and universities to choose from. The problem of choosing a university and study program is complicated by the phenomenon of unique combination of each university and study program. Approximate number of universities could be obtained from Ref. [1]: in EU - 3400; in Europe - 5700; in the World – 31100.

The school graduates do not have information for assessment of their individual opportunities for successful completion of various study programs at various universities. It is leading to the problems of high dropout rates [2-3], students switching from one study program to another [4], increasing the period of obtaining university education [5] and low efficiency of university education [6-8]. At the same time, universities already possess such information, but do not share it with other stakeholders of educational market. Elimination of data silo may create new services for data-driven educational decisions and rise the efficiency of university education [9-10].

The R&D Project: "Advanced data analysis and forecasting in education" [11] is dedicated to solving this problem. The project aims to create pan-European undergraduate study program search system (USPSS) for school graduates. Analysing information from Lithuanian universities and electronic school diaries on individual performance of pupils and students, the system advises the most appropriate study programs to school graduates. Analysis of data on VGTU students confirmed the existence of central tendency, a positive correlation between the level of performance in certain school disciplines and the results of further education in certain university study programs [12]. Accordingly, it is possible to create predictive models of university learning results based on demographic data and school learning results. The USPSS suggestions are based on criteria of successful completion: a) low probability of dropout, failing exams or passing after many retakes, switching study programme; b) high probability of admission, successful graduation, and career etc.

This work is aimed to solve the following tasks.

1. Review the existing literature on the application of Data Science in the education sector to predict students' achievements and learning results.
2. Analyse the student data of Lithuanian universities, including students' school marks and their adjustment into a single scale in accordance with the LAMA BPO algorithm, to identify the correlation between students' characteristics at the time of their admission and their future performance at university.
3. Examine LAMA BPO algorithm, based on students' data from two Lithuanian universities and confirm or refute the possibility of creating an optimal algorithm for adjusting school grades into a single scale based on a positive relationship between the level of performance in certain school disciplines and the results of further education in certain study programs at the university.
4. Suggest further research directions.

1. Literature review

Many authors tested different Data Science methods to predict students' achievements and learning results. Existing research can be divided based on data sources used: 1) behavioural data (clickstream analytics, pages visited, time spent, etc.) from online learning systems [13-16]. 2) student's previous grades - articles based on student performance should be considered in more detail, since they have a similar data structure and algorithms.

Alsuwaiket et al. [17] applied *Naive Bayes* and *Random Forest* algorithms to forecast students' second year averages based on their first-year averages. The results obtained by researchers especially interesting as data gathered for research was from similar study programs and departments with similar names as we are going to analyse in case of VGTU (Vilnius Gediminas Technical University) (study programmes *Civil Engineering*, *Computer Science*, *Electrical and Computer Systems*, *Engineering*, *Mathematics*, *Mechanical Engineering*, *Business*). *Random Forest* algorithm was more accurate. Accuracy measured by using Area Under the Curve (AUC) method, that provides an aggregate measure of forecasting model performance.

Similar to previous research Hasan et al. [18] used classification approach to forecast future university learning outcomes, based on relatively small dataset of 1170 students. They applied *K-Nearest Neighbors* and *Decision Tree* (with *ID3* algorithm to determine the root). *Decision Tree* was more accurate in predicting student performance.

Kostopoulos et al. [19] applied *Bayesian network*, *Naïve Bayes*, *Decision tree* (C4.5), *K-Nearest Neighbors* and *Sequential Minimal Optimization* to forecast future marks based on dataset of 340 students. Marks of two previous academic semesters were used as independent variables. The highest accuracy is 72.94% was achieved by the model developed based on *Naïve Bayes*.

As Panessai et al. [20] aimed their study to predict the student's attrition and failing to complete the course, it unites them with this research work. Other similarity is study programme name they analysed: *Software Engineering* at Universiti Pendidikan Sultan Idris. In present research we are going to analyse 595 student's data from VGTU *Software Engineering* programme (Panessai et al. [20] analyzed data of 123 students) along with student's data from other 42 VGTU and 8 LSU study programs. But Panessai et al. [20] used learning outcomes data of already admitted students, grades for: mid-term exam, group project, quizzes, coursework, and other assessments.

Contrary we use school learning outcomes as independent variables to predict student's failing to complete the course and other student's performance results. Further add to differences is classification approach chosen by Panessai et al. [20]. They applied *Naïve Bayes*, *Generalized Linear classification*, and *Decision Tree* methods to predict the final grade (A, B, C, D, E, F). The highest accuracy (79.18%) was achieved by forecasting model developed based on *Decision Tree* (C4.5 algorithm) method.

Li et al. [21] clustered student grades across three college courses (*Physics*, *Career Planning and Management*, *Chinese Language and Literature*) for teachers to make better decisions on delivering quality education. This is an example of unsupervised learning and therefore has no target variable to predict. By applying *Fuzzy C-means* method Li et al. [21] divided learning outcomes of three university courses in four groups: "Great", "Good", "Average" and "Bad". Therefore, the applicability of research results is limited to visualization and analysis of the status quo. The analysis is based on a small dataset of 246 students from Huaqiao University. In conclusion authors propose possibility of application of *Support Vector Machines* and *Artificial Neural Network* to predict students' grades.

Rajab and Ramadan [22] aimed their research at predicting *Grade Point Average* (GPA) based on grades in the first and second semester, non-public and social factors such as living location (dormitory or apartment), and attendance. But instead of employing regression-based forecasting methods for predicting numeric values, researchers used classification methods dividing predicted GPA into four classes: First Class, Upper Class, Lower Class, Pass. As the research is based on a small dataset of 72 student records from the College of Health Science at the State University of Zanzibar, the results obtained may be just random coincidences. For example, only 2 out of 72 rows in the dataset classified as "Pass" (the lowest result), both rows represented female students with very good attendance and good/very good level of secondary school completion, living in apartments, etc. Based on such characteristics, the Data Science model may predict similar female students to get the lowest results. However, for male students, such a forecast is impossible, since there are no corresponding records in the data based on which the forecasting model was trained.

The choice of modeling methods used by researchers looks controversial since all methods are based on a *Decision Tree* algorithm: *C4.5*, *Iterative Dichotomiser 3* (ID3), *Classification and Regression Tree* (CART), and *Chi-square automatic interaction detection* (CHAID). For example, it is not clear why they use ID3 (precursor to C4.5 algorithm) together with its successor C4.5. Predictably, ID3 was least correct, while CART demonstrated an exceptional accuracy of 100%.

Table 1 represents the applicability of university learning results prediction studies to current research.

2. Methodology

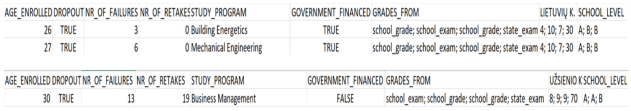
2.1. Data preparation: analysis, parsing, transformation, and cleaning

Current research based on the data of 24341 Vilnius Gediminas Technical University (VGTU) and 2137 Lithuanian Sports University (LSU) students. Characteristics and structure of the original dataset contain following statements.

1. Admission year to university Bachelors' study program in (VGTU: from 2009 to 2021, LSU from 2010 to 2019).
2. Country where admitted student studied at school (19 counties).
3. Gender (M/F).
4. Year of school graduation (from 1979 to 2021).
5. Age at the time of the admission to university (from 17 to 58).
6. Drop out of the university ("true" - student dropped-out, "false" - not dropped-out, "changed in the study programme" - student switch to another study program at the same university).
7. Number of courses failed to complete (study courses with the mark lower than 5 in a 10-point system was identified as failed). In the dataset minimum number of failed courses is 0 maximum is 49 of failed courses per student.
8. Number of retakes. In the dataset minimum number of retakes is 0 maximum is 94 per student.
9. Average university grade.
10. Name of study program.
11. Year of university graduation.
12. Type of financing (state funded position / paid by student).
13. Information about school subjects (courses) and school marks includes:
    i) course name;
    ii) level of education: "A", "B", "S", "M", "T";
    iii) type of grade: "annual school grade", "school exam" or "state maturity exam";
    iv) course mark.

Table 1. Applicability of university learning results prediction studies to current research.

Type of ML	Aim	Target variable	More accurate method/ other methods	Accuracy, %	Independent variables	Dataset size - Number of students	Study programme	Refs.	Applicability
SL	Not specified by authors	Second year average, classified by groups (Fail, First, Lower second, Pass, Third, Upper second)	*Random Forest/ Naive Bayes*	94.2	Course marks (module, coursework, exam), other course-related, study program-related attributes	230 823	Civil Engineering, Computer Science, Electrical and Computer Systems, Engineering, Mathematics, Mechanical Engineering, Business	[17]	Very limited by di erences in: discretization of target variable, and data structure
SL	Admission and scholarships decisions	GPA classified by groups (fail, fair, good, very good)	*Decision Tree/Bayesian Network*	86	Grades in previous years by semesters, non-public and social factors such as Family Job, Ethnic, Religion	20492	Not specified by the authors	[23]	Limited by di erences in: discretization of target variable, and data structure
SL	Decision of additional attention from teacher	GPA classified by groups (First Class, Upper Class, Lower Class, Pass)	*CART/ C4.5*, *Decision Tree (C4.5)*, *Decision Tree (ID3)*, *CHAID*	100	Grades in first and second semester, non-public and social factors such as living location (dormitory or apartment), and attendance	72	Healthcare	[22]	Questionable due to di erences in: research aims, discretization of the target variable, student's field of study, volume, and the structure of the data used for forecasting
SL	Decision of additional attention from teacher	Course final examination mark (numbers classified as groups)	*Decision Tree* (ID3)/ *K-Nearest Neighbors*	94.88	Course name, test mark, attendance mark, presentation mark, assignment mark, midterm mark, final examination mark	1170	Not specified by the authors	[18]	Very limited by di erences in: research aims, discretization of target variables, and the structure of the data used for forecasting
SL	Decision of additional attention from teacher	Course final grade classified by groups (Poor, Good, Very good, Excellent)	*Naïve Bayes/Bayesian network*, *Decision tree* (C4.5), *K-Nearest Neighbors*, *Sequential Minimal Optimization*	72.94	Two semester course marks (oral, tests, exam, overall)	340	Not specified by the authors, since the dataset is taken from open sources	[19]	Very limited by di erences in: research aims, discretization of target variable, and the structure of the data used for forecasting
SL	Decision of additional attention from teacher	Final grade classified by groups (A, B, C, D, E, F)	*Decision Tree (C4.5)/ Naïve Bayes*, *Generalized Linear classification*	79.18	Grades: mid-term exam, group project, quizzes, coursework, other assessments	123	Software Engineering	[20]	Aim of the study partly coincides with the aim of current research, but applicability of study is limited by di erences in discretization of target variable, and the structure of the data used for forecasting
USL	Clustering and visualizing student grades across three courses to make better decisions on delivering quality education	Final grades of three college courses clustered by groups (Great, Good, Average, Bad)	*Fuzzy C-means clustering*	Not used in USL	Grades of three college courses (College Physics, Career Planning and Management for College Students, College Chinese Language and Literature)	246	Information Technology	[21]	Not possible due to a fundamental di erence in the research aim



AGE	ENROLLED	DROPOUT	NR_OF_FAILURES	NR_OF_RETAKES	STUDY_PROGRAM	GOVERNMENT_FINANCED	GRADES_FROM	LIETUVIŲ K. SCHOOL_LEVEL
26	TRUE		3	0	Building Energetics	TRUE	school_grade; school_exam; school_grade; state_exam	4; 10; 7; 30 A; B; B
27	TRUE		6	0	Mechanical Engineering	TRUE	school_grade; school_exam; school_grade; state_exam	4; 10; 7; 30 A; B; B

AGE	ENROLLED	DROPOUT	NR_OF_FAILURES	NR_OF_RETAKES	STUDY_PROGRAM	GOVERNMENT_FINANCED	GRADES_FROM	UŽSIENIO K SCHOOL_LEVEL
30	TRUE		13	19	Business Management	FALSE	school_exam; school_grade; school_grade; state_exam	8; 9; 9; 70 A; A; B

Fig. 1. Fragment of the dataset with school marks data.

Fig. 1 shows different types, levels, and scales of school knowledge assessments for courses in "Lithuanian" and "Foreign language". Here we see a difference in the grades of the same school subject. The first student's annual school grades in Lithuanian are 4 (A level) and 7 (B level). In addition, the student has 10 (B level) for school exam and 30 for state exam. Also, we can see another student with double annual school marks in foreign language: 9 (A level) and 9 (B level), and 8 for the school exam (A level) with 70 for the state exam (Fig. 1).

To build correctly functioning forecasting models and predict successful completion of study programme, it is required to adjust all school marks into a single scale. The algorithm developed by Lithuanian Association of Higher Education Institutions for General Admission (LAMA BPO) was chosen for that purpose [24] . LAMA BPO is an Association the purpose of which is to organize and coordinate general admission to Lithuanian institutions implementing study and training programs. This algorithm was implemented in the processing of the data of Lithuanian universities and schools. According to the LAMA BPO algorithm, school final exams and annual grades have several types of adjusting coefficients.

The state exam grades of 16-100 points scale are adjusted into the single scale according to the Eq.(1):

$$Y = 4 + (X - 16) * 0.07143 \quad (1)$$

where Y represents the grade adjusted to the single scale; X represents the grade of the state exam. As can be seen from Table 2, the value of annual school grades is significantly inferior to the value of final school exam grades. According to the algorithm, if there are several grades for a school discipline, then after adjusting them into a single scale, the maximum is selected. This grade is considered an assessment of the pupil's knowledge of the school course. Following the LAMA BPO algorithm, let's determine the grade in Lithuanian based on the data from Figure 1: the annual school grade of A-level will become 3.6 (see Table 2), the annual school grade of B-level will become 2.3, the school exam grade of B-level will become 7.2, and the state exam grade will become 5 - see Eq.(1). The maximum value is 7.2, this value is taken as an assessment of Lithuanian.

To process data and convert school grades into university entry points, based on the LAMA BPO algorithm, KNIME Analytics Platform was used (Fig. 2, 3, 5). Since in the original dataset, all grades of the applicant are in one cell through the separating ";" (Fig. 1), then to process each grade, program divided each cell by number of grades. For this, purpose node "cell Splitter" is used (Fig. 2). Because each grade of school discipline has three types of

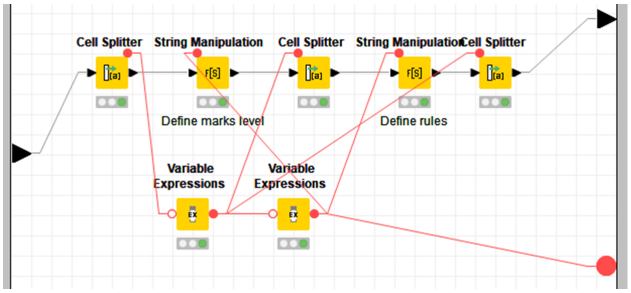


Fig. 2. Converting school marks into a single scale: data preprocessing.

characteristics: grade value, grade type and level, then to process, for example, four different grades in Lithuanian at the end of this step we get 12 columns with grades characteristics. Variable Expressions - assigns the iteration number to the name of each column so that the grades and their characteristics are not confused (Fig. 2).

The program shown in Fig. 3 creates rules to convert school marks by using the LAMA BPO algorithm. Each rule is written by two nodes ("Table Creator" and "String Manipulation") then rules are merged in one table (Fig. 3). Fig. 4 shows conversion table with the rules (conditions) and conversion values.

The last phase of converting school marks into a single scale is implementation of rules from Fig. 4 and Eq.(1) - as presented in Fig. 5. At the final stage of data processing, cleaning was implemented to process missing values, missing students, etc. This form initial dataset for creation of prediction models of successful completion of study programs.

2.2. Modelling the academic performance at university

For each of the mentioned criteria of successful completion each study program of each university, it is required to create a predictive model. For that purpose, forecasting factory was developed based on KNIME Analytics Platform. forecasting models were built using data of 24341 Vilnius Gediminas Technical University (VGTU) and 2137 Lithuanian Sports University (LSU) students who studied from 2010 to 2020. At the top level of the loop the forecasting factory iterates over universities, picking students data of new university at each iteration.

The second level of the loop iterates over parameters of successful completion of study programme (GPA, number of fails to complete the courses, etc.). The third level of the loop iterates over study programmes creating predictive models to forecast parameters of successful completion for each study program. For example, the forecasting model to predict the number of fails to complete courses in the study programme "Air Traffic Control" was developed based on *Ensemble of Regression Trees* algorithm (Fig. 6).

Table 2. Adjusted school annual and final exam grades according to the LAMA BPO algorithm.

Assessment scales	School ten-point scale						
	4	5	6	7	8	9	10
Adjusted A-level, S-level, no-level school exam grade	4,4	5,3	6,1	7,0	7,9	8,7	9,6
Adjusted B-level school exam grade	3,3	4,0	4,6	5,3	5,9	6,5	7,2
Adjusted A-level, S-level, no-level annual school grade	3,6	3,9	4,2	4,5	4,8	5,1	5,4
Adjusted B-level annual school grade	1,8	2,0	2,1	2,3	2,4	2,6	2,7

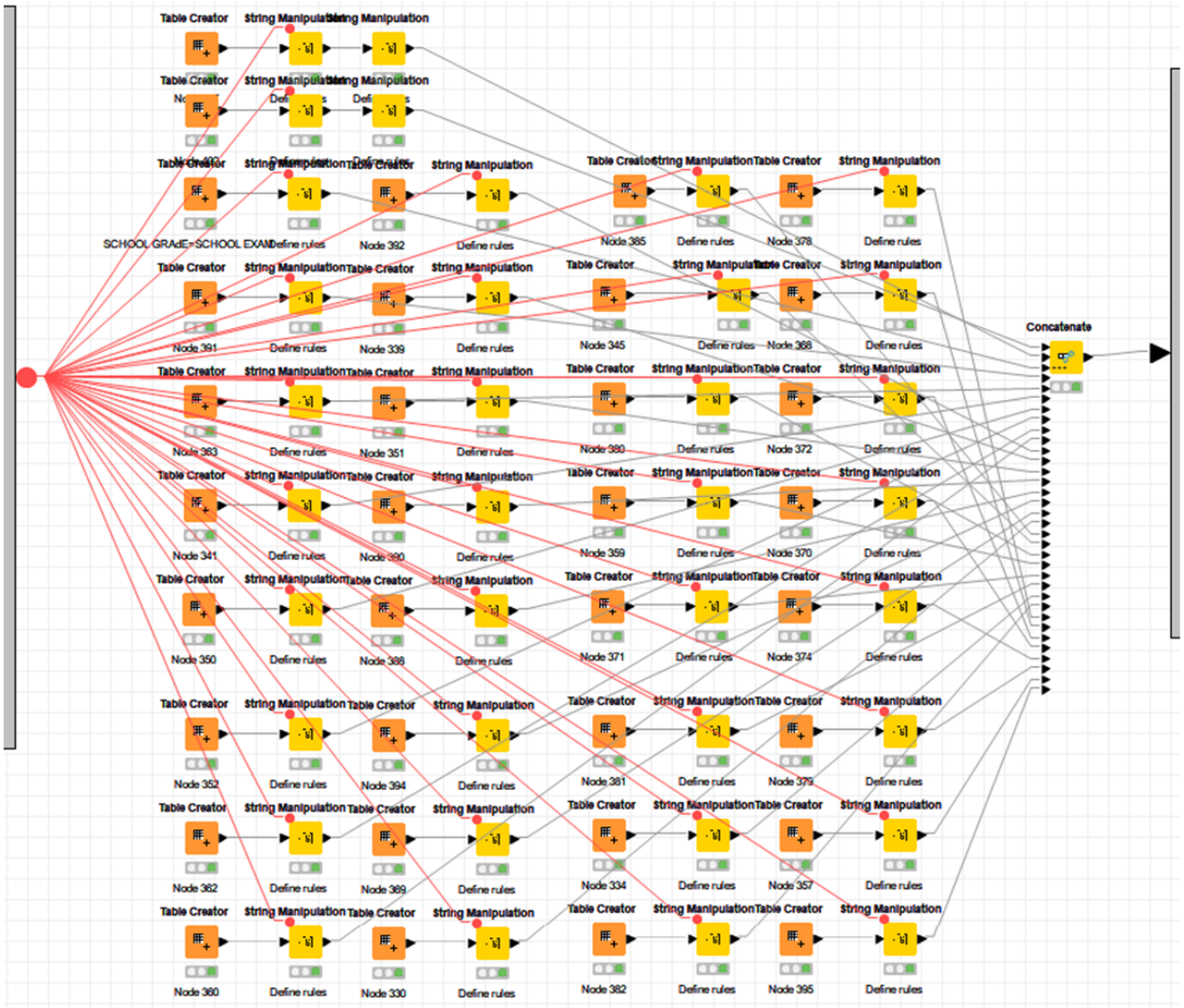


Fig. 3. Converting school marks into a single scale: formation of rules for adjusting school grades into a single scale.

RULE	Result
MISSING $LIETUVIU K._Arr[0]$	$LIETUVIU K._Arr[...
$GRADES_FROM_0_Arr[0]$="state_exam"	$LIETUVIU K._Arr[...
$LIETUVIU K._Arr[0]$>=2.49 AND $LIETUVIU K._Arr[0]$<3.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	3.3
$LIETUVIU K._Arr[0]$>=3.49 AND $LIETUVIU K._Arr[0]$<4.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	3.6
$LIETUVIU K._Arr[0]$>=4.49 AND $LIETUVIU K._Arr[0]$<5.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	3.9
$LIETUVIU K._Arr[0]$>=5.49 AND $LIETUVIU K._Arr[0]$<6.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	4.2
$LIETUVIU K._Arr[0]$>=6.49 AND $LIETUVIU K._Arr[0]$<7.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	4.5
$LIETUVIU K._Arr[0]$>=7.49 AND $LIETUVIU K._Arr[0]$<8.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	4.8
$LIETUVIU K._Arr[0]$>=8.49 AND $LIETUVIU K._Arr[0]$<9.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	5.1
$LIETUVIU K._Arr[0]$>=9.49 AND $LIETUVIU K._Arr[0]$="school_grade" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	5.4
$LIETUVIU K._Arr[0]$>=2.49 AND $LIETUVIU K._Arr[0]$<3.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	1.5
$LIETUVIU K._Arr[0]$>=3.49 AND $LIETUVIU K._Arr[0]$<4.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	1.8
$LIETUVIU K._Arr[0]$>=4.49 AND $LIETUVIU K._Arr[0]$<5.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.0
$LIETUVIU K._Arr[0]$>=5.49 AND $LIETUVIU K._Arr[0]$<6.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.1
$LIETUVIU K._Arr[0]$>=6.49 AND $LIETUVIU K._Arr[0]$<7.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.3
$LIETUVIU K._Arr[0]$>=7.49 AND $LIETUVIU K._Arr[0]$<8.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.4
$LIETUVIU K._Arr[0]$>=8.49 AND $LIETUVIU K._Arr[0]$<9.49 AND $GRADES_FROM_0_Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.6
$LIETUVIU K._Arr[0]$>=9.49 AND $LIETUVIU K._Arr[0]$="school_grade" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.7
$LIETUVIU K._Arr[0]$>=2.49 AND $LIETUVIU K._Arr[0]$<3.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	2.5
$LIETUVIU K._Arr[0]$>=3.49 AND $LIETUVIU K._Arr[0]$<4.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	3.3
$LIETUVIU K._Arr[0]$>=4.49 AND $LIETUVIU K._Arr[0]$<5.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	4.0
$LIETUVIU K._Arr[0]$>=5.49 AND $LIETUVIU K._Arr[0]$<6.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	4.6
$LIETUVIU K._Arr[0]$>=6.49 AND $LIETUVIU K._Arr[0]$<7.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	5.3
$LIETUVIU K._Arr[0]$>=7.49 AND $LIETUVIU K._Arr[0]$<8.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	5.9
$LIETUVIU K._Arr[0]$>=8.49 AND $LIETUVIU K._Arr[0]$<9.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	6.5
$LIETUVIU K._Arr[0]$>=9.49 AND $LIETUVIU K._Arr[0]$="school_exam" AND $SCHOOL_LEVEL_0_Arr[0]$="B"	7.2
$LIETUVIU K._Arr[0]$>=2.49 AND $LIETUVIU K._Arr[0]$<3.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	3.3
$LIETUVIU K._Arr[0]$>=3.49 AND $LIETUVIU K._Arr[0]$<4.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	4.4
$LIETUVIU K._Arr[0]$>=4.49 AND $LIETUVIU K._Arr[0]$<5.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	5.3
$LIETUVIU K._Arr[0]$>=5.49 AND $LIETUVIU K._Arr[0]$<6.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	6.1
$LIETUVIU K._Arr[0]$>=6.49 AND $LIETUVIU K._Arr[0]$<7.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	7.0
$LIETUVIU K._Arr[0]$>=7.49 AND $LIETUVIU K._Arr[0]$<8.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	7.9
$LIETUVIU K._Arr[0]$>=8.49 AND $LIETUVIU K._Arr[0]$<9.49 AND $GRADES_FROM_0_Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	8.7
$LIETUVIU K._Arr[0]$>=9.49 AND $LIETUVIU K._Arr[0]$="school_exam" AND ($SCHOOL_LEVEL_0_Arr[0]$="A" XOR $SCHOOL_LEVEL_0_Arr[0]$="S" XOR $SCHOOL_LEVEL_0_Arr[0]$="M" XOR MISSING $SCHOOL_LEVEL_0_Arr[0]$)	9.6

Fig. 4. Conversion rules (based on LAMA BPO algorithm).

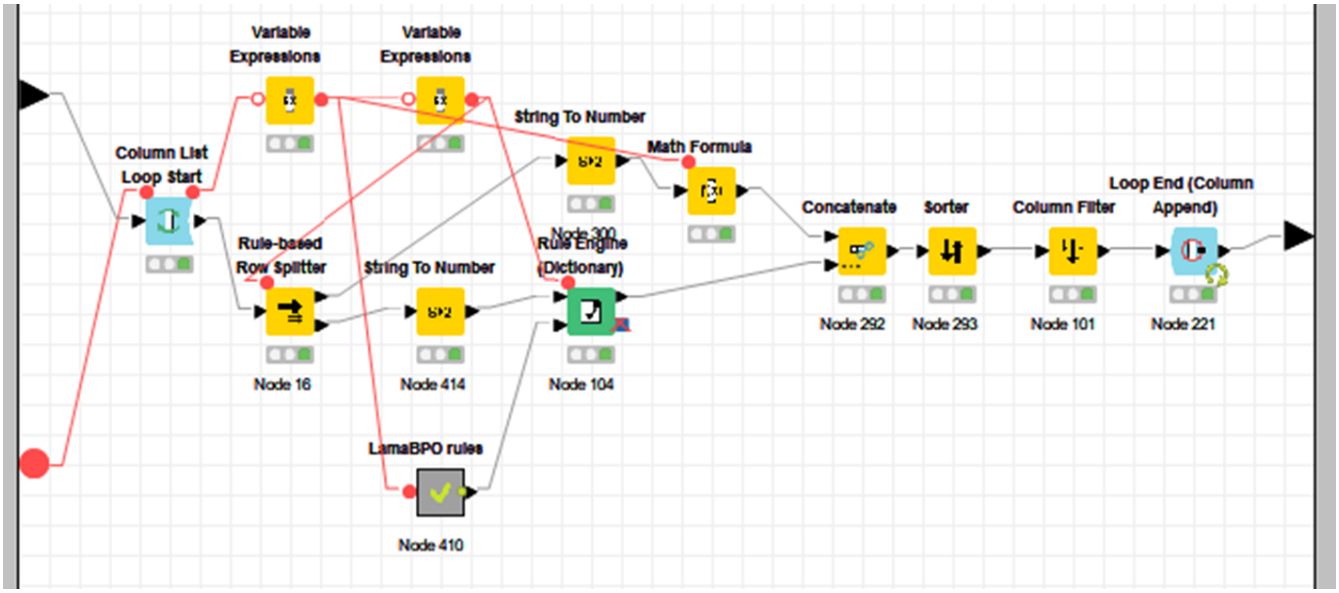


Fig. 5. Converting school marks into a single scale: adjusting school grades into a single scale.

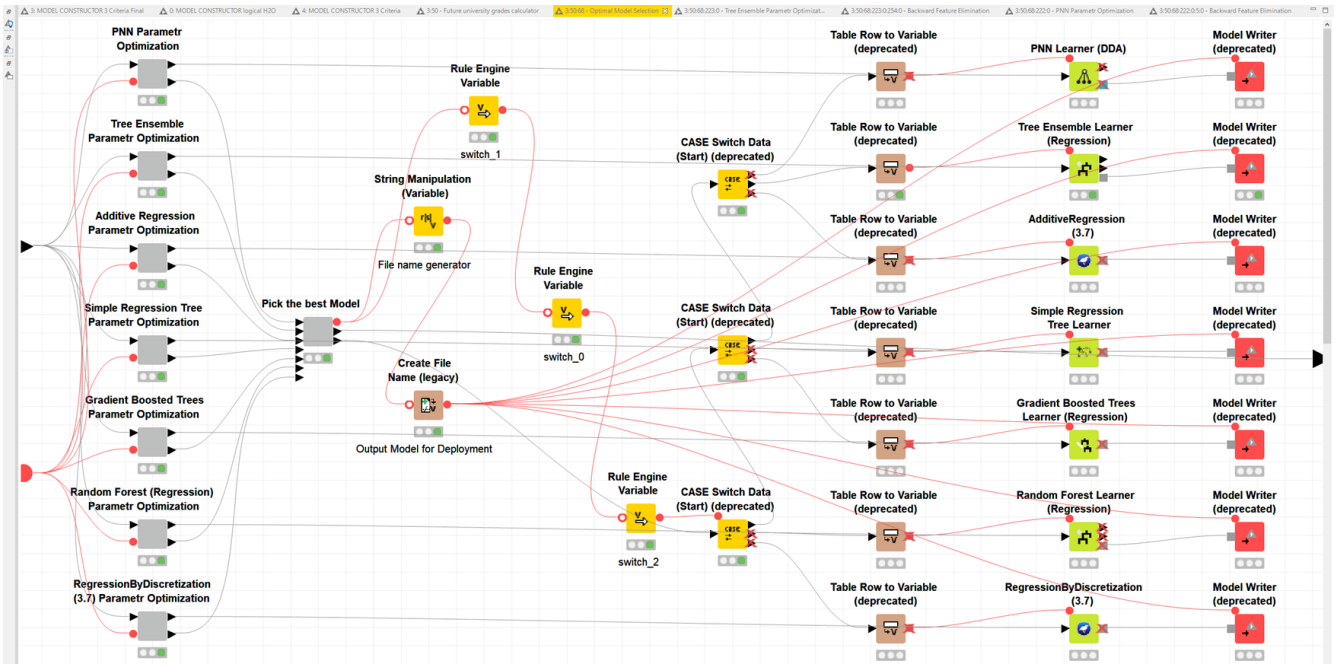


Fig. 6. Predictive model creation to forecast the number of fails to complete courses in the study programme "Air Traffic Control" based on *KNIME Analytics* platform.

For each parameter of successful completion of study programme, models are created and tested, based on seven basic forecasting algorithms (those algorithms differ depending on predicted parameter). The models go through stages of parameter optimization and backward feature elimination to improve their accuracy. As a result, the model with the smallest prediction error is selected to be stored in USPSS.

Iterative work to improve the accuracy of the models created by forecasting factory led to the result after which any of forecasting factory setting updates caused a decrease in the accuracy of the models created. Since the potential to increase the accuracy of forecasting models by further program changes has been depleted, the decision was made to try changes in pre-processing logic of incoming data.

This led the authors of the article to formulate a hypothesis: the LAMA BPO algorithm is non-optimal in adjusting final school grades of Lithuanian school graduates into a single scale. Probable reason is that LAMA BPO algorithm overestimates the level of school knowledge of some pupils and underestimates it for others.

This can reduce university performance, because in some cases, "weaker" pupils are admitted to university instead of more "powerful" ones.

As an optimality criterion for the new algorithm, we choose the minimization of the error of the predictive models of the results of subsequent education at the university.

This criterion assumes that there is a positive relationship between the level of performance in certain school disciplines and the results of further education in certain university study programmes (simplified - the higher the performance at school, the more successful the university education will be). This relation was confirmed based on the analysis of data of VGTU students [12].

It follows that the more accurately we determine school performance, the more accurately we can predict university performance (the error of the predictive model will be less).

To prove the hypothesis, it is enough to find at least one option of adjusting school graduate's grades into a single scale, different from the LAMA BPO algorithm, which will provide consistently more accurate predictions of further learning results at the university. To do this, different options were tested to bring the school grades into a single scale. The generation of models for predicting results of education at the university was carried out, the accuracy parameters of the models were calculated, and compared.

The root-mean-square error (RMSE) was used as a criterion for the assessment of accuracy of the models. RMSE is quadratic mean of differences between predicted values and observed values [25]. It shows how concentrated the predicted and actual values of indicators are [26]. The lower the RMSE of the model, the more accurate the forecast it gives [27].

The difference from the LAMA BPO algorithm was that not only the maximum values of the grades adjusted into a single scale were considered. The decision was made to create predictive models based on minimal, mean, median and max marks for each school course:

1) MINIMUM – from all marks for each school discipline, a minimum is selected;
2) MEAN - the average is calculated for each discipline;
3) MEDIAN - the "middle" of a sorted list of marks;
4) MAX - from all marks for each school discipline, a maximum is selected.

All 4 variants alternative to the LAMA BPO algorithm were developed based on the principle of increasing the weight of annual school grades, with the weight of school and state exams unchanged. The variant that provided the highest accuracy of the forecast, i.e., the smallest RMSE is presented in Table 3.

Accuracy parameters of predictive models (RMSE) are saved for subsequent analysis. Those RMSE of models predicting GPA are presented in Table 4. The method of constructing the most accurate forecasting model is indicated by following abbreviations.

1. ERT, *Ensemble of Regression Trees*, described by Loh [29];
2. BART, *Bayesian Additive Regression Trees*, described by Friedman [30];
3. SGBRT, *Stochastic Gradient Boosted Regression Trees*, described by Friedman [31];
4. RF, *Random Forest Regression*, described by Breiman [32];
5. PNN, *Probabilistic Neural Network*, described by Berthold & Diamond [33].

The table cells with the RMSE of most accurate forecasting models are highlighted in colour (Table 4).

3. Results and discussion

It was found that the RMSE of models created based on data adjusted by LAMA BPO algorithm is higher than the similar indicator of the models based on data adjusted to single scale based on algorithm prepared by the authors [28], see Table 4.

Based on Table 4, we can conclude the following outcomes.

1. For the deployment of the models, it is better to choose from models developed on different data sources (Min, Median, Mean, Max mark), e.g., for some study programs higher accuracy could be achieved by considering not Median or Mean mark, but min marks as for study programme *Multimedia Design*.
2. The most accurately determining the level of knowledge of school disciplines, and therefore allowing to predict the success of studying more accurately at the university, are the median and mean of marks values of the grades adjusted to a single scale.
3. The average RMSE of models built based on data adjusted by the LAMA BPO algorithm is the highest and is 0.574 (Table 4). The average RMSE of models built based on data adjusted by the algorithm proposed by the authors of the article is significantly lower. It should be emphasized that our algorithm is more flexible and the system each time selects and saves a model built on such type of data (min, median, mean, max), that allows us most accurately capture the dependence of learning results at the university on learning results at school. In most cases, these are models built on median or mean grades, but sometimes the minimum or maximum grades allow to predict future learning outcomes more accurately.
4. The standardization based on the LAMA BPO algorithm provides a higher error, regardless of which type of the school mark values is considered (min, median, mean, max).
5. The lower RMSE (about 0.08) resulted from the *Probabilistic Neural Network* model based on data of minimum marks of school disciplines. Model was built based on data on 574 students and tested based on data on 143 students (80% and 20% of the initial data on 717 students of the study programme *Multimedia Design*). It indicates the successful determination of central tendency in the data and good forecasting results of the model.

Table 3. Adjusted school annual and final exam grades (prepared by the authors).

Assessment scales	School ten-point scale						
	4	5	6	7	8	9	10
Adjusted A-level, S-level, no-level school exam grade	4,4	5,3	6,1	7,0	7,9	8,7	9,6
Adjusted B-level school exam grade	3,3	4,0	4,6	5,3	5,9	6,5	7,2
Adjusted A-level, S-level, no-level annual school grade	4,0	4,8	6,0	7,0	7,7	8,5	9,4
Adjusted B-level annual school grade	3,1	4,0	4,5	5,2	5,7	6,4	7,1

Table 4. RMSE of models predicting GPA for VGTU study programmes

NR	Study programmes	N_{ST}	Method of adjusting school grades into a single scale								DSA
			LAMA BPO algorithm				Novel algorithm prepared by authors [28]				
			Min mark	Median mark	Mean mark	Max mark	Min mark	Median mark	Mean mark	Max mark	
1.	Air Traffic Control	99	0.429	0.425	0.424	0.444	0.409	0.399	0.399	0.428	RF
2.	Aircraft Piloting	139	0.536	0.488	0.518	0.529	0.468	0.452	0.452	0.478	SGBRT
3.	Architecture	658	0.607	0.567	0.584	0.591	0.556	0.562	0.568	0.567	ERT
4.	Aviation Mechanics Engineering	326	0.659	0.565	0.570	0.580	0.562	0.552	0.553	0.569	SGBRT
5.	Avionics	344	0.798	0.728	0.722	0.792	0.706	0.692	0.692	0.697	ERT
7.	Building Energetics	315	0.673	0.641	0.641	0.663	0.618	0.608	0.612	0.616	BART
8.	Business Logistics	211	0.560	0.583	0.586	0.557	0.525	0.532	0.538	0.509	SGBRT
9.	Business Management	649	0.592	0.585	0.583	0.664	0.548	0.547	0.547	0.548	PNN
10.	Civil Engineering	912	0.682	0.589	0.595	0.594	0.579	0.577	0.577	0.580	RF
11.	Creative Industries	818	0.565	0.554	0.564	0.609	0.549	0.532	0.526	0.537	BART
12.	Multimedia Design	717	0.743	0.521	0.578	0.690	0.081	0.14	0.14	0.14	PNN
...	...										
43.	Transport Engineering Economics and Logistics	825	0.732	0.799	0.734	0.733	0.711	0.705	0.702	0.621	RF
	Average		0.538	0.523	0.531	0.574	0.449	0.431	0.428	0.436	

N_{ST} Number of students (2010-2020)
DSA Data Science Algorithm for predictive model construction (EW, Mean mark)

The experiment of changing the algorithm of adjusting school marks into a single scale allowed to achieve a significant reduction in RMSE for different VGTU study programs, for example:

a) from 0.690 (LAMA BPO algorithm) to 0.081 (author's algorithm [28]) – study programme *Multimedia Design*.
b) from 0.758 (LAMA BPO algorithm) to 0.216 (author's algorithm [28])) - study programme *Security Systems Engineering*.

This proves research hypothesis that LAMA BPO algorithm is non-optimal in assessing the school performance, it shows the lower accuracy of models based on central tendency: positive correlation between the level of performance in certain school disciplines and the results of further education in certain university study programs [12]. Probable reason is that LAMA BPO algorithm overestimates the level of knowledge for some pupils and underestimates it for others. This can reduce university performance, because in some cases, "weaker" pupils are admitted to university instead of more "powerful" ones.

On the other hand, the author's method [28] of adjusting school grades into a single scale is not optimal too. It should be considered as the first step of Bayesian optimization [34] in which from 100 to 1000 different adjusting algorithms (tables with values of grades adjusted to single scale) will be used to generate data and build forecasting models, assess their accuracy, and guess directions to construct new promising adjusting algorithms, based on the *Tree-structured Parzen Estimator* approach. New promising adjusting algorithms are then evaluated (to minimize RMSE of forecasting models) and search for new promising adjusting algorithms continued. The result will be the optimal adjusting algorithm.

Combining the data on school and university performance from the main Lithuanian universities, we will get a representative database, therefore will be able to develop the optimal algorithm for adjusting school marks into a single scale.

Conclusions

Several important conclusions can be drawn from this study.

1. The research hypothesis was proven. LAMA BPO algorithm is non-optimal in assessing the school performance and could be improved. Several predictive models have been created according to various methods of adjusting school grades into a single scale. As a result, it was proved that the LAMA BPO algorithm is not the most accurate when it comes to determining the level of knowledge of Lithuanian pupils.
2. The research aim was achieved: a system was developed to forecast a school graduate's university results with a minimum forecasting error.
3. A new hypothesis was put forward: by combining data on school and university performance from the main universities of Lithuania, it is possible to construct a system for calculating the parameters of the optimal algorithm by the criteria of minimizing the errors of forecasting models: successful studies at the university (higher GPA); dropout; not passing exams; retakes; change study program, unemployment and employment prospects, etc.
4. Novel algorithm [28] could be recommended for use by LAMA BPO.

Acknowledgements

Great thanks to Lithuanian Business Support Agency (VšĮ Lietuvos verslo paramos agentūra) for funding R&D project: "Advanced data analysis and forecasting in education" (J05-LVPA-K-04-0132, "Pažangi duomenų analizė ir prognozavimas švietimo sričiai"). The presented results were by-product of aforementioned R&D project.

Authors' contributions

Aleksei Iurasov conceived and designed the analysis, reviewed the existing literature, wrote the introduction, literature review and conclusions of the article. Artur Iurasov collected the data, performed the analysis, wrote the methodology, results, and discussion of the article. All authors read and approved the final manuscript.

Conflicts of interest

There are no conflicts to declare.

Abbreviations

AI	-	Artificial Intelligence
AUC	-	Area Under the Curve
BART	-	Bayesian Additive Regression Trees [30]
ERT	-	Ensemble of Regression Trees [29]
GPA	-	Grade Point Average
KNIME	-	Konstanz Information Miner
LAMA BPO	-	Lietuvos Aukštųjų Mokyklų Asociacija Bendrajam Priėmimui Organizuoti (lith), Association of the General Admission to Lithuanian Universities
LSU	-	Lithuanian Sport University
ML	-	Machine Learning
PNN	-	Probabilistic Neural Network [33]
R&D	-	Research and Development
RF	-	Random Forest Regression [32]
RMSE	-	Root Mean Square Error
SGBRT	-	Stochastic Gradient Boosted Regression Trees [31]
SL	-	Supervised learning
USL	-	Supervised learning
USPSS	-	Undergraduate Study Program Search System
VGTU	-	Vilnius Gediminas Technical University

References

1. Webometrics Ranking of World Universities (2022). Countries arranged by Number of Universities in Top Ranks [online]. http://www.webometrics.info/en/node/54 (Accessed 23 July 2022).
2. Heublein U. (2014) Student Drop-out from German Higher Education Institutions. – *European Journal of Education* Vol. 49 No. 4, pp. 497-513. https://doi.org/10.1111/ejed.12097
3. Hess F. (2018) The college dropout problem [online] https://www.forbes.com/sites/frederickhess/2018/06/06/the-college-dropout-problem/ (Accessed 23 July 2022).
4. Willis M. (2005) Why do students switch from one university to another: The view of students studying for a foreign degree in Hong Kong. – *Journal of Marketing for Higher Education* 15(1), 23-49.
5. Salas-Velasco M. (2007) The transition from higher education to employment in Europe: the analysis of the time to obtain the first job. – *Higher Education* 54(3), 333-360.
6. Salas-Velasco M. (2020) The technical efficiency performance of the higher education systems based on data envelopment analysis with an illustration for the Spanish case. – *Educ. Res. Policy Pract.* 19(2), 159–180
7. Abdalmenem S. A., Owda R. O., Al-Hila A. A., Abu Naser S. S., Al Shobaki M. J. (2018) The Performance Efficiency of University Education between Reality and Expectations. – *International Journal of Academic Management Science Research (IJAMSR)* 2(10), 66-76.
8. Johnes J., Portela M., Thanassoulis E. (2017) Efficiency in education. – *Journal of the Operational Research Society* 68(4), 331-338.
9. Lim L.-A., Gasevic D., Matcha W., Ahmad Uzir N., Dawson S. Impact of learning analytics feedback on self-regulated learning: Triangulating behavioural logs with students' recall – In: 11th International Conference on Learning Analytics and Knowledge: The Impact we Make: The Contributions of Learning Analytics to Learning, LAK 2021; Virtual, Online; United States; 12 April 2021 through 16 April 2021; Code 168184. P. 364-374.
10. Rossen A., Boll C., Wolf A. (2019) Patterns of overeducation in Europe: The role of field of study. IZA J. – *Labor Policy* 9(3), 1–48
11. R&D Project: "Pažangi duomenų analizė ir prognozavimas švietimo sričiai" (Advanced data analysis and forecasting in education, Project number: J05-LVPA-K-04-0132).
12. Iurasov A. (2021) New e-business model: undergraduate study program search system. – *International Journal of Learning and Change* Available at: http://dx.doi.org/10.1504/ijlc.2021.10035252
13. Feng M., Beck J., Heffernan N., Koedinger K. (2008) Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standarized Test?.In Baker and Beck (Eds.) – In: Proceedings of the 1st international conference on educational data mining (pp. 107–116). Montreal.
14. Puška A., Puška E., Dragić L., Maksimović A., Osmanović N. (2021) Students' satisfaction with E-learning platforms in Bosnia and Herzegovina. – *Technology, Knowledge and Learning* 26(1), 173-191. doi: 10.1007/s10758-020-09446-6
15. Nouri J., Saqr M., Fors U. (2019) Predicting performance of students in a flipped classroom using machine learning: towards automated data-driven formative feedback. – In: 10th International conference on education, training and informatics (ICETI 2019).
16. Poloju K. K., Naidu V. R. (2020) Impact of E-tools in Teaching and Learning for Undergraduate Students. – In: Innovations in Electronics and Communication Engineering (pp. 783-790). Springer, Singapore.
17. Alsuwaiket Mohammed, Blasi Anas and Al-Tarawneh Khawla (2020) Refining Student Marks based on Enrolled Modules' Assessment Methods using Data Mining Techniques. – *Engineering, Technology and Applied Science Research* Vol. 10, pp. 5205-5210.
18. Hasan H.M., Rabby Akm Shahariar Azad, Islam Mohammad, Hossain Syed (2019) Machine Learning Algorithm for Student's Performance Prediction. – In: 10th International Conference on Computing, Communication and Networking Technologies. Kanpur, India.
19. Kostopoulos G., Livieris I. E., Kotsiantis S. and Tampakas V. (2017) Enhancing high school students' performance based on semi-supervised methods. – In: 2017 8th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1-6.
20. Panessai I.Y., Lakulu M.M., Rahman M.H., Noor N.A., Salleh N.S., Bilong A.A. (2019) PSAP: Improving Accuracy of Students' Final Grade Prediction using ID3 and C4.5. – *International journal of artificial intelligence* 6, pp. 125-133.
21. Li Y., Gou J., and Fan Z. (2019) Educational data mining for students' performance based on fuzzy C-means clustering. – *The Journal of Engineering* 2019(11), pp. 8245-8250.
22. Rajab A.M., Ramadhan R.M. (2019) Application of Data Mining Techniques in Students' Performance Prediction and Analysis. – *International Journal of Academic Information Systems Research* 3, 1–9.
23. Janecek P. (2007) A comparative analysis of techniques for predicting academic performance. – In: 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports.
24. Order of the Minister of Education, Science and Sports Nr. V-718, May 5 2022. – https://www.e-tar.lt/portal/lt/legalAct/d7d19ff0cc4211ec8d9390588bf2de65.
25. Wang W., Lu Y. (2018) Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. – In: IOP conference series: materials science and engineering (Vol. 324, No. 1, p. 012049). IOP Publishing.
26. Chai T., Draxler R. R. (2014) Root mean square error (RMSE) or mean absolute error (MAE). – *Geoscientific Model Development Discussions* 7(1), 1525-1534.
27. Hodson T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. – *Geoscientific Model Development* 15(14), 5481-5487.
28. Aleksei Iurasov (2022) HELM-SCM, v.1. – API
29. Loh W.-Y. (2014) Fifty Years of Classification and Regression Trees. – *International Statistical Review* 82(3) – https://doi.org/10.1111/insr.12016

30. Friedman Jerome H. (2002) Stochastic gradient boosting. – *Computational Statistics and Data Analysis* 38(4), 367–378.
31. Friedman Jerome H. (2001) Greedy function approximation: A gradient boosting machine. – *The Annals of Statistics* 29(5) – https://doi.org/10.1214/aos/1013203451
32. Breiman, L. (2001). Random forests. – *Machine Learning* 45(1) – https://doi.org/10.1023/A:1010933404324
33. Berthold M. R., Diamond J. (1998) Constructive training of probabilistic neural networks. – *Neurocomputing* 19(1–3) – https://doi.org/10.1016/S0925-2312(97)00063-5
34. Mockus J., Eddy W., Reklaitis G. (2013) Bayesian Heuristic approach to discrete and global optimization: Algorithms, visualization, software, and applications (Vol. 17). – Springer Science & Business Media.