Assessment of Students' Achievements and Competencies in Mathematics Using CART and CART Ensembles and Bagging with Combined Model Improvement by MARS

Snezhana Gocheva-Ilieva * , **Hristina Kulina** and **Atanas Ivanov**

Department of Mathematical Analysis, University of Plovdiv Paisii Hilendarski, 4000 Plovdiv, Bulgaria;
kulina@uni-plovdiv.bg (H.K.); aivanov@uni-plovdiv.bg (A.I.)
* Correspondence: snow@uni-plovdiv.bg

Abstract: The aim of this study is to evaluate students' achievements in mathematics using three machine learning regression methods: classification and regression trees (CART), CART ensembles and bagging (CART-EB) and multivariate adaptive regression splines (MARS). A novel ensemble methodology is proposed based on the combination of CART and CART-EB models in a new ensemble to regress the actual data using MARS. Results of a final exam test, control and home assignments, and other learning activities to assess students' knowledge and competencies in applied mathematics are examined. The exam test combines problems on elements of mathematical analysis, statistics and a small practical project. The project is the new competence-oriented element, which requires students to formulate problems themselves, to choose different solutions and to use or not use specialized software. Initially, empirical data are statistically modeled using six CART and six CART-EB competing models. The models achieve a goodness-of-fit up to 96% to actual data. The impact of the examined factors on the students' success at the final exam is determined. Using the best of these models and proposed novel ensemble procedure, final MARS models are built that outperform the other models for predicting the achievements of students in applied mathematics.

Keywords: mathematical competency; assessment; machine learning; classification and regression tree; CART ensembles and bagging; ensemble model; multivariate adaptive regression splines; cross-validation

1. Introduction

The quality of mathematics training in higher education is essential for competitive future professional achievements of students in engineering, software, economics and other specialties. Alongside traditional teaching and learning methods in mathematics, increasingly various information technologies, computer- and mobile-based methods are applied using specialized software, as well as methodologies that involve project and team work, group discussions, role playing, blended learning etc. [1,2]. In particular, in the last two decades, an overall vision for teaching mathematical subjects in connection with their possible practical applications has been actively developed regarding the concept of competence. The concept is defined in [3] as follows: "Mathematical competency is understood as the ability to understand, judge, do, and use mathematics in a variety of intra- and extra-mathematical contexts and situations in which mathematics plays or could play a role."

In the context of higher education in engineering [3,4], the following eight key mathematical competencies are formulated: C1—thinking mathematically; C2—reasoning mathematically; C3—posing and solving mathematical problems; C4—modeling mathematically; C5—representing mathematical entities; C6—handling mathematical symbols and formalism; C7—communicating in, with, and about mathematics; C8—making use of aids and tools. Based on these competencies, specific teaching and learning methods for mathemat-

ics and assessment of student knowledge in engineering higher education and assessment standards were discussed in [5,6], which report the results of statistical modeling of data from a summative and competency-based assessment test using two machine learning and data mining techniques—cluster analysis and classification and regression trees (CART). With the aid of these methods, models are obtained for classification and determination of dependencies, for predicting student achievements based on the grades from a linear algebra and analytical geometry test and a short 10 min general mathematical competence test. A recent paper [7] analyzes the results from a written test of knowledge and the accompanying self-assessment survey of the examined students for individual problems using the CART method.

To improve and measure the level of student knowledge and competencies, in literature both traditional and modern cognitive and statistical approaches are applied. Standard multivariate statistical methods for the assessment of student knowledge are used, for example, in [8] to measure mathematical competencies of students upon admission to university with the help of Rash analysis and other analyses to provide insights into the measures' reliability and validity. A methodology for improving communication competencies and skills by learning mathematics in engineering degree specialties is presented with examples in [9]. Numerous publications use educational data mining (EDM) to establish classifications and dependencies in heterogeneous types of information related to education at all levels. EDM encompasses several research fields, such as data mining, machine learning (ML) and statistics. A recent review article [10] provides systematic information and analysis of a large number of studies, which use soft computing methods in EDM and ML for 2010–2018. The authors emphasize that decision tree, random forest, artificial neural network (ANN), fuzzy logic, support vector machine (SVM) and genetic/evolutionary algorithms are a few examples of soft computing approaches that, given enough data, can successfully deal with uncertainty, qualitatively stated problems and incomplete, imprecise or even contradictory data sets. These types of methods have a wide scope of application for research into various problems. Classifying and predicting students' academic success are carried out in [11] using several decision tree algorithms. A model is obtained which successfully predicts 79% of the grades of the students involved. To predict the student's performance in the Introduction to Informatics module, the authors in [12] applied six ML techniques, namely naïve Bayes, decision trees C4.5, NN, instance-based learning, logistic regression and SVM. It was found that the naïve Bayes algorithm is the most appropriate technique. In [13], decision trees, artificial neural networks and naïve Bayes models are built to predict students' academic performance based on their academic record, personal data and social information. Decision tree classification and regression models are built and studied for evaluation of mathematical competencies and student success in [7], achieving model performance of over 90% for both types of models. The highly effective data mining and ML technique, random forest (RF), is used in [14] for predicting students' dropout from university. In [15], a comparative study of seven predictive models for high school student performance in mathematics is performed using ML, deep learning and other techniques. The RF models show the best qualities with over 90% predictive performance. The same authors apply four ML techniques in [16] and also build hybrid models utilizing principal component analysis. The best predictive results of up to 98% with minimal relative error are obtained by RF models. SVM, ANN, fuzzy functions and other types of ML models are obtained and analyzed in several papers [17–21]. Other examples of ML methods with applications in education can be found in review papers [22–24]. Recent advances related to all kinds of ensemble learning algorithms, frameworks and methodologies, and their applications, can be found in [25].

The aim of this study is to demonstrate a combined traditional and competence-oriented approach to conducting an examination test in mathematics, as well as to determine factors affecting students' mathematics achievements and competencies using powerful predictive ML techniques. A case study is performed with results from the final exam in the course of Applied Mathematics with the first year students in specialty Busi-

ness Information Technologies at University of Plovdiv Paisii Hilendarski, Bulgaria, which also includes as its principal component a small practical project. The main predictors used in the analyses are students' grades from ongoing testing during the trimester (control works and home assignments), attendance at lectures and laboratory practice, as well as the scores on individual problems in the exam and a small practical project. The modeling of the empirical data is performed with the methods CART and CART ensembles and bagging (CART-EB). To improve the result of the prediction of the exam test points, the best of these models are assembled with MARS.

This is the first time that the CART-EB method is applied for statistical modeling of data in the field of education. Another contribution is the use of the MARS method to generate new ensemble models from other ensemble models.

2. Materials and Methods

2.1. Methodology

The main part of each training process in education is the assessment of knowledge and skills, acquired by the students at a given stage. Depending on the curriculum for a given mathematical subject, in order to pass the exam, the student attends a certain number of lectures and laboratory practices, takes intermediate tests (control tests), solves assignments at home, works on individual or group projects, prepares presentations etc. Usually this type of control is evaluated with a certain score. This combination of activities is denoted as preparatory. At the end of their education, the students take a final exam, which can be written, oral or a combination of the two, or another type of assessment. It is assumed that the grade from the exam is influenced by the combination of preparatory activities during the course of the education. In order to apply a competency-oriented approach to the assessment of acquired knowledge and skills, a small practical-oriented project is used as a component of the final exam. All preparatory activities and the components of the exam test can be assigned a certain type of measurement and the respective dataset can be derived, where the grades are presented as variables. Exam grades can be considered to be a dependent or target variable, and the rest are predictors. Potential predictors are, for example, homework grades, course project grades, reports, gender of the student, the high school he or she graduated from.

Our experimental empirical study sets out to perform the following tasks:

- Construction of the integrated competency-based test for the final exam in mathematics;
- Construction, analysis and improvement of predictive models for evaluation of students' achievements using ML techniques;
- Application of the models for determining the importance of the preparatory activities and individual components of the exam to its assessment and, in particular, the importance of the project.

In essence, these tasks point to finding hidden similarities and patterns in the data using ML regression-type modeling techniques.

2.2. Machine Learning Methods Used for Statistical Analyses

The term ML (also referred to as learning analytics) denotes a class of methods and algorithms of artificial intelligence. Usually ML is used for classification and regression problems, and self-learning is achieved through various algorithms for cross-validation, improvement of model accuracy and fitting quality. This is achieved by combining features of computational statistics tools, numerical methods, optimization methods, probability theory, graph theory etc. ML methods are nonparametric and allow the detection of nonlinearities and relationships in the data without the need to model them explicitly; that is, they are data driven. Their core advantage is the generation of numerous distribution free and robust models, among which the most adequate and optimal model in a given sense can be selected. The following ML methods are widely used to model educational data: logistic regression, cluster analysis, decision trees (CART), support vector machines (SVM),

multivariate adaptive regression splines (MARS), random forest (RF), neural networks (NN), fuzzy logic and others [10,24].

2.2.1. Classification and Regression Tree (CART)

The CART method [26] is a typical representative of decision tree algorithms and can be used for classification or regression. The main concept of the method is to classify the data from the training dataset through a recursive procedure into a binary tree structure with nodes. At each stage, the cases in the current node, called parent node, can be split into two child nodes according to the threshold value of some predictor variable. The predicted value in a terminal node is simply the average of the response values located in that node [24]. The threshold value is determined by a greedy algorithm, which checks all variables and their values so that the model minimizes the current selected type of summary error of the predicted values or other criteria in the terminal nodes of the tree. The splitting criterion for regression trees can be least squares or least absolute deviation. Once the tree is constructed, branches that do not contribute to the improvement of the model are removed and a final pruned tree is obtained. The researcher presets the settings and hyperparameters to select an optimal model, the type of cross-validation or other ML procedure, and adjusts the algorithm. For more details, see [27,28].

2.2.2. CART Ensembles and Bagging (CART-EB)

There are cases in which regression CART models may show instability in prediction under the influence of outliers, unsignificant predictors, predictors with small variation and others. There may also arise overfitting of the model [29]. Then, it is appropriate to use an ensemble of trees in combination with a bagging (also known as bootstrap aggregation) algorithm. There are many ML methods involving these techniques. In the current study, the algorithm of the CART-EB method was applied using the analysis engine CART ensembles and bagger of the Salford Predictive Modeler software suite [30]. Some other implementations in literature can be found in [31,32]. The initial CART tree of the ensemble is constructed with the entire data sample and all predictors. Then, it is pruned using 10-fold cross-validation. Bagged trees are built independently one from the other on bootstrap samples with or without repeated cases. They use a random subset of predictor variables at each decision split as in the RF algorithm. Ensemble trees can be built as exploratory (unpruned) maximal trees or they can be pruned by cross-validation. The case-predicted value is the average of the predictions of all the trees in the ensemble.

2.2.3. Multivariate Adaptive Regression Splines (MARS)

MARS is a nonparametric data mining and machine learning method, developed in [33]. If the dependent variable (here *Exam*) is $y = y(X)$ and $X = (X_1, X_2, \ldots, X_p)$ are p predictors with dimension n, the regression MARS model $\hat{y} = \hat{y}_{[M]}$ has the following form:

$$\hat{y}_{[M]} = b_0 + \sum_{j=1}^{M} b_j BF_j(X) \tag{1}$$

where $b_0, b_j, j = 1, 2, \ldots, M$ are the coefficients in the model, $BF_j(X)$ are its basis functions (BF), M is the number of BFs. The one-dimensional BF is written in the form

$$BF_j(X) = \max_{X_k}(0, X_k - c_{k,j}) \text{ or } BF_j(X) == \max_{X_k}(c_{k,j} - X_k, 0), \tag{2}$$

where the nodes $c_{k,j} \in X_k$ are determined by the MARS algorithm. For the nonlinear interactions, BFs are built as products of other BFs.

The control parameters chosen by the researcher are the maximum number of basis functions and the maximum number of their multipliers (i.e., degree of interactions) in BFs. The algorithm involves two steps. The first step starts by setting b_0 (for example, $b_0 = \min_{1 \le i \le n} y_i$) and then the model is complemented consistently by BFs of type (2). For each

model with a given number of BFs, the MARS algorithm defines variables and nodes so as to minimize a predefined loss function, such as the root mean square error. In the second step, BFs that do not contribute significantly to the accuracy of the model are removed. For more details, see [33].

2.2.4. Model Evaluation Metrics

In this study, the best ML models were selected using the highest coefficient of determination R^2 and the minimum values of the root mean square error (RMSE) given by the expressions

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

where $\hat{y}_i$ and y_i stand for model predicted and *Exam* values, respectively.

The performance of the models was also evaluated using the Theil's forecast accuracy coefficient U_{II} [34]:

$$U_{II} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2}}. \tag{4}$$

The lower the value of the coefficient, the better the accuracy of the model. The coefficient U_{II} is dimensionless and is used to compare models obtained by different methods, as well as to identify large values. The model is considered to be of good quality when (4) is less than 1.

When choosing from nested models, the parsimony principle was applied [35].

3. Results

3.1. Test Design

An experiment was conducted with the Applied Mathematics course. The final exam test combines three main components with problems in mathematical analysis, probability theory and applied statistics. It includes:

- Problems in math analysis (5 problems),   15 points,   50%;
- Problems in probability (2 problems),   5 points,   17%;
- A small practical project in applied statistics,   10 points,   33%.

The percentage indicates the relative weight within the total number of 30 points for the entire exam. Unsolved problems are evaluated with 0 points. A sample version of the exam test with 7 type variations is given in Figure 1. Each student works on an individual test. It needs to be noted that the problems in the first two components are of traditional type; these problems have been used in exams in this course of studies over the last 7–8 years. The added project includes some general instructions without explicitly stating how the problem is to be solved.

The exam was taken by 68 first-year students in the specialty of Business Information Technologies at the Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria. According to the first trimester curriculum, these students have taken a linear algebra and analytic geometry course, and during the second trimester, the course in Information Technology for Mathematics, where students are trained to use Wolfram Mathematica to solve mathematical problems using a computer. The current course in Applied Mathematics is in the third trimester.

The results of the preliminary activities and the final exam in number of points are described with the variables: *Exam* (total exam points, up to 30), *Math_An* (mathematical analysis, up to 15), *Stat* (statistics, up to 5), *Project* (up to 10), *A1_12* (home assignment 1, up to 12), *A2_20* (assignment 2, up to 20), *CW1_30* (homework 1, up to 30), *CW2_30* (homework

2, up to 30), *Attn_Lect* (attendance to lectures, up to 10) and *Attn_Labs* (attendance to labs, up to 10).

Exam in Applied Mathematics for the specialty Business Information Technologies, 1ˢᵗ year

Substitute: *p* = penultimate digit of faculty number +1; *q* = last digit of faculty number +1

Problem 1 (2 + 2 + 3 + 3 + 5 points). The function $f(x) = \dfrac{px + q}{x^2 - 9}$ is given.

A) Determine the definition domain of $f(x)$.

B) Find the first derivative of $f(x)$.

C) According to the sign of the first derivative to find the intervals of increase and decrease of the function. You can draw the necessary graphs.

D) Determine the limits $\lim\limits_{x \to -3} f(x)$ and $\lim\limits_{x \to 3} f(x)$.

E) Calculate the area locked between the function and the abscissa axis for $x \in [0, 1]$.

Problem 2 (3 + 2 points). Gymnast trains a combination with bats. In the basket are collected 6 pairs of new and 5 pairs of old bats. If she plays with new bats, the probability of dropping a bat is $p / 20$, and if she plays with old bats, the probability is $(p + q) / 20$.

A) Find the probability of dropping a bat if the gymnast chooses any pair of bats from the basket.

B) Find the probability of not dropping a bat if she has chosen a new pair.

Problem 3 (Small Project, 10 points). Load the Real_estate3.sav file with SPSS. Delete the first 5 $(p + q)$ lines. The file contains data on houses for sale in a large city. The variables are:

No - number of the case, X1 transaction date, X2 house age, X3 distance to the nearest metro station, X4 number of convenience stores, X5 latitude, X6 longitude, Y house price of unit area.

A) Calculate the descriptive statistics of the variables. Check the type of distribution.

B) Use appropriate statistical analyses to investigate whether there is a significant dependence in the dataset.

C) Remove the insignificant variables and repeat the analysis. Build the model to identify the factors that affect the price and interpret the result.

D) What percentage of the data explains the model?

The decision should be explained on a separate sheet.

Total: maximum 30 points.

Send the solutions with the Wolfram Mathematica and SPSS working files, to the teacher's e-mail.

Figure 1. Example of the exam test in Applied Mathematics.

3.2. Measurement of Competencies by the Exam Test

Following the recommendations of [4], the experience in [5,36] and with the aid of a three-dimensional scale, we defined the correspondence between the eight competencies and the elements of the exam test, as shown in Table 1. Here T1–T5 mean subproblems A–E in problem 1; S1 and S2 in problem 2; P1–P4, the instructions to the project. It is shown that all competencies are included with the exception of C7 because the exam is individual and does not allow for any communication with other students and/or external sources. In addition, Figure 1 shows that the probability theory problem *Stat* requires a solution with pen and paper and does not duplicate the project. As a whole, the project is independent in terms of curriculum covered and supplements the competencies, which are not included in the first two test components. The level of solution of the project indicates the degree to which the students have acquired the necessary knowledge and skills in statistics in order to solve on their own a complete mathematical problem—from the data, through the analyses to the interpretation of the results obtained.

It should be noted that the students have solved the project in different ways, with different methods. Some managed to make only descriptive statistics, with different statistics selected. Other students continued with cluster analysis, factor analysis or principal component analysis. More often, regression analysis was performed in one-dimensional or multidimensional case.

Table 1. Assessments of the level of mathematical competencies in problems from the *Exam* test [1].

Competency		*Exam* Elements										
		T1	T2	T3	T4	T5	S1	S2	P1	P2	P3	P4
C1	Thinking mathematically	−	−	−	−	0	+	+	0	0	0	0
C2	Reasoning mathematically	−	−	0	−	0	0	0	−	0	0	−
C3	Problem solving	0	0	0	0	+	+	+	0	0	0	−
C4	Modeling mathematically	−	−	−	−	−	−	−	+	+	+	0
C5	Representation	0	0	+	+	0	+	−	−	−	−	−
C6	Symbols and formalism	0	0	0	0	0	−	−	0	0	0	−
C7	Communication	−	−	−	−	−	−	−	−	−	−	−
C8	Aids and tools	−	0	−	+	+	−	−	+	+	+	−

[1] The meaning of the signs: +, "very important"; 0, "medium important"; −, "less important".

3.3. Initial Processing and Analysis of the Data

Descriptive statistics of the variables used in the study are given in Table 2 and the distributions in the form of box plots with unstandardized data are shown in Figure 2a,b. Table 2 and Figure 2b show that the mean values of the results for *Stat* and *Project* are quite low, and their median is 0. The reason is because only 31 students, or 45%, worked on the project. In addition, Table 2 and Figure 2a,b lead us to the conclusion that most of the variables are not normally distributed (*A2_20*, *CW_30*, *Stat*, *Project* etc.). This is evidenced by the relatively high values of the ratios of skewness/std. error of skewness, kurtosis/std. error of kurtosis, as well as from the box plots. For example, for the target variable *Exam* we have the ratios $\frac{Skewness}{Std.\ Err.\ Skewneess} = \frac{1.14}{0.29} = 3.931 > 1.96$ and $\frac{Kurtosis}{Std.\ Err.\ Kurtosis} = \frac{1.85}{0.57} = 2.643 > 1.96$, which is an indication for non-normal distribution of the variable. In addition, a one-sample Kolmogorov–Smirnov test with Lilliefors significance correction was applied, which reaffirms that *Exam* did not follow a normal distribution as the calculated *p*-value is 0.000. Furthermore, the relationships between the variables are hidden and possibly highly nonlinear.

Table 2. Descriptive statistics of the initial predictors and the target variable.

| Statistics	*Attn_Lect*	*Attn_Labs*	*A1_12*	*A2_20*	*CW1_30*	*CW2_30*	*Math_An*	*Stat*	*Project*	*Exam*
Mean	6.87	5.63	6.65	12.46	10.07	7.88	8.67	1.76	2.48	12.91
Median	8.00	5.00	7.00	15.75	9.25	6.00	8.75	0.00	0.00	12.00
Std. Deviation	3.44	3.11	3.71	6.94	8.35	7.24	3.64	2.234	3.18	4.88
Variance	11.82	9.67	13.75	48.13	69.73	52.46	13.25	4.99	10.12	23.78
Skewness	−0.81	0.19	−0.32	−0.90	0.51	0.44	0.12	0.81	0.86	1.14
Std. Error of Skewness	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
Kurtosis	−0.72	−1.38	−0.85	−0.79	−0.68	−1.12	−1.018	−0.91	−0.79	1.85
Std. Error of Kurtosis	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Range	10	10	12	19.5	30.0	22	2.0	0.00	10.0	25.0
Minimum	0	0	0	0.0	0.0	0	15.0	7.00	0.0	4.0
Maximum	10	10	12	19.5	30	22	8.7	1.76	10.0	29.0

3.4. Results from the CART Models

Multiple regression trees were built using the CART method. The dependent variable is *Exam* and the factors on which its values depend are the remaining nine variables, i.e., *Math_An*, *Stat*, *Project*, *A1_12*, *A2_20*, *CW1_30*, *CW2_30*, *Attn_Lect*, and *Attn_Labs*. The objective was to define which independent factors have the strongest influence on the *Exam* and to what extent. Before applying the algorithm, hyperparameters m1 (minimum cases in parent node) and m2 (minimum cases in child node) were set. Regression tree procedure on the learn (training) set is carried out using 10-fold cross-validation, which

is recommended for small samples [27,28]. The least squares method was selected as a splitting criterion.
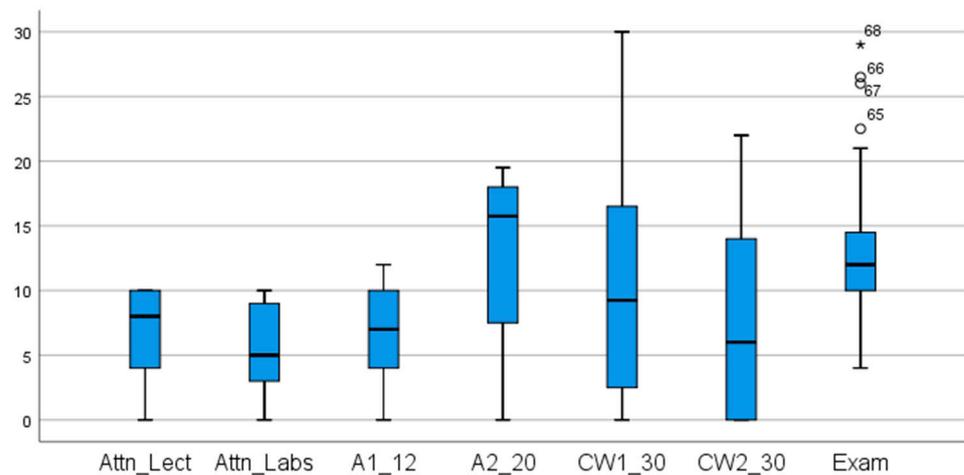

(a)
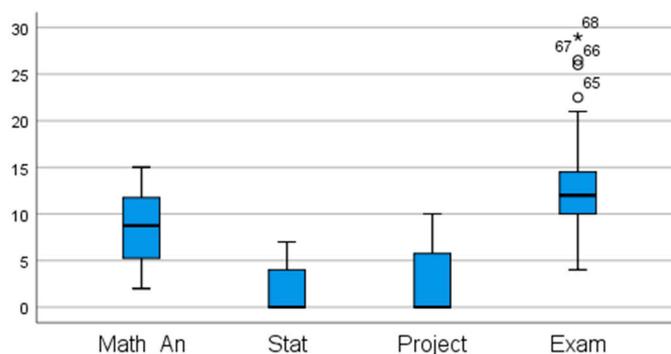
(b)

Figure 2. Box plots of the initial predictors and target variable *Exam*, used in the statistical analyses (nonstandardized): (**a**) preliminary activities and *Exam*, (**b**) *Exam* elements.

For m1 = 5, m2 = 2 in Figure 3a, a diagram is shown of the relative error of the generated CART models depending on the number of their terminal nodes. For the case of m1 = 6, m2 = 3, the scheme with the relative errors of the constructed models is presented in Figure 3b. Relative errors are calculated as the ratio of the least square error of the current model divided by the root node error. Models with relative errors distinguished by one standard error (1 SE) are colored in green. This means that all models in green from Figure 3a,b can be considered as a set of competing models. From Figure 3a, it is evident that the model with a minimum relative error of 0.321 has 13 terminal nodes. We denote it by M_1. In addition, two models were analyzed—the M_2 model with a minimum number of 9 nodes and the maximum M_3 model with 22 nodes. Besides these, in the same way we denote the model with a minimum relative error of 0.310 and 11 terminal nodes with M_4, the model with 9 terminal nodes with M_5, and the model with 19 terminal nodes with M_6, respectively.

Table 3 contains summary statistics for the competing six CART models M_1, M_2, ... , M_6 that are selected. We compare the two optimal models M_1 and M_4. Although model M_4 has larger constraints of m1 = 6 and m2 = 3, it shows the highest value of R^2 test = 0.698, and the minimum value of RMSE Test = 2.694. At the same time, this model is inferior to the prediction statistics, especially with the relatively large RMSE = 1.517, which is 16% higher than that of M_1. From the first group of "finer" models, M_1 is comparable to

M_4 with R^2 Test = 0.621 compared to 0.698 for M_4 (1%), RMSE Test = 2.743, with a small difference of 0.049, or less than 2%. The goodness-of-fit R^2 Learn = 0.928 of the M_1 model is 3% higher than the M_4 statistic (0.902). Next, we make a comparison between M_1 and M_6. The statistics of these two models are almost identical, but the M_6 model is more complex as it contains 19 terminal nodes compared to the 13 of M_1. From the set of competing models considered, we should reject M_2 and M_5. This is due to the most unsatisfactory statistics—the smallest R^2 and the largest RMSE Learn of the prediction. Model M_3 has a less favorable relative error compared to M_1 (with 4%), outperforming M_1 narrowly for the main indicators (4) by 1 to 9%. Since the M_3 model is the most complex, having 22 terminal nodes, compared to the 13 of M_1, we have to apply the parsimony principle [34,35] (see also Figure 3a). We will further consider CART models M_1 and M_4. Note that all Theil's coefficients are sufficiently small.
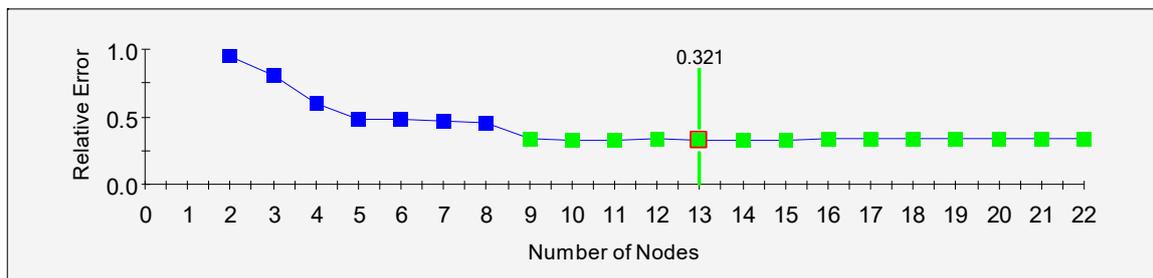
(**a**)
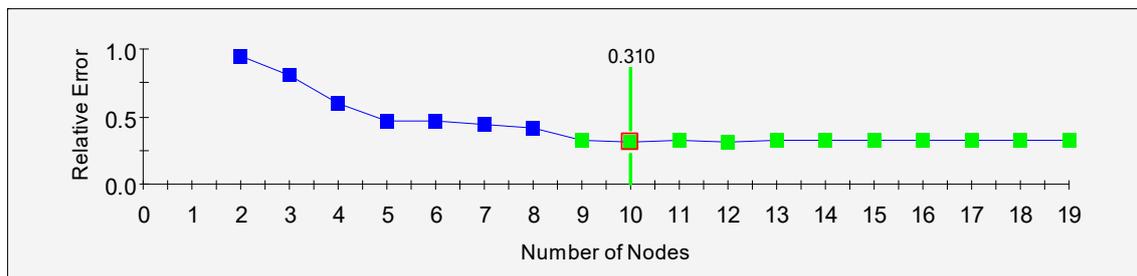
(**b**)

Figure 3. Relative error of the constructed classification and regression tree (CART) models depending on the number of terminal nodes: (**a**) m1 = 5, m2 = 2; (**b**) m1 = 6, m2 = 3.

Table 3. Summary statistics of the selected regression CART models for assessment of students' achievements.

Statistic	Model					
	M$_1$	**M$_2$**	**M$_3$**	**M$_4$**	**M$_5$**	**M$_6$**
Terminal nodes	13	9	22	10	9	19
m1-m2	5-2	5-2	5-2	6-3	6-3	6-3
Relative error	0.321	0.333	0.335	0.310	0.319	0.325
R^2, Test	0.621	0.683	0.678	0.698	0.690	0.685
R^2, Learn	0.928	0.892	0.940	0.902	0.899	0.929
RMSE, Test	2.743	-	-	2.694	-	-
RMSE, Learn	1.298	1.588	1.188	1.517	1.616	1.291
Theil's U_{II}	0.0089	0.0133	0.0074	0.0121	0.0137	0.0088

Table 4 presents the values of the relative importance of the nine factors studied on the *Exam* points according to their participation in the exploratory Learn CART trees. Here, too, stability is clearly visible, with small differences. The largest factor of importance

(100 relative points) is the *Project* factor—the points from the exam obtained for solving the small practical project. The next main factors, in descending order of their relative importance, are the scores on *Math_An* (93–95 relative units) and *A2_30* (50–54 units). The points from solving the problems in the *Stat* statistics have a small impact, within 20–22 relative units. Table 4 also shows that the influence of the predictors obtained in the chosen optimal model M_1 is almost identical to that of model M_3, as well as the other maximal model M_6.

Table 4. Relative importance of the initial predictors used in the selected CART models.

Predictors	CART Models					
	M_1	M_2	M_3	M_4	M_5	M_6
Project	100	100	100	100	100	100
Math_An	95.25	93.76	95.27	93.25	93.08	94.15
A2_20	54.34	50.52	54.59	52.80	50.52	54.17
CW1_30	37.79	35.99	38.63	36.36	35.99	37.97
A1_12	36.02	35.10	36.29	35.29	35.29	36.68
Attn_Labs	34.15	34.27	34.45	34.27	34.27	33.90
CW2_30	22.66	19.72	23.31	22.12	19.72	22.84
Stat	21.74	20.28	21.57	22.04	20.28	21.58
Attn_Lect	6.23	5.03	6.90	6.26	5.03	6.39

The calculation of the coefficients of importance in Table 3 is performed using sequential aggregation. At level 0, the mean value of target (*Exam*) as predicted by the model is calculated for the entire sample and the RMSE is calculated. At the first split (as shown in Figure 4), the CART algorithm selected the *Math_An* as a splitter predictor and its threshold value is 11.25. After the split, the predictions (mean values) are calculated in both child nodes along with their RMSEs. The relative improvement of the current accuracy achieved is calculated against the root, and the value obtained is added to the coefficient of importance of the predictor. The process is repeated until the tree is built.
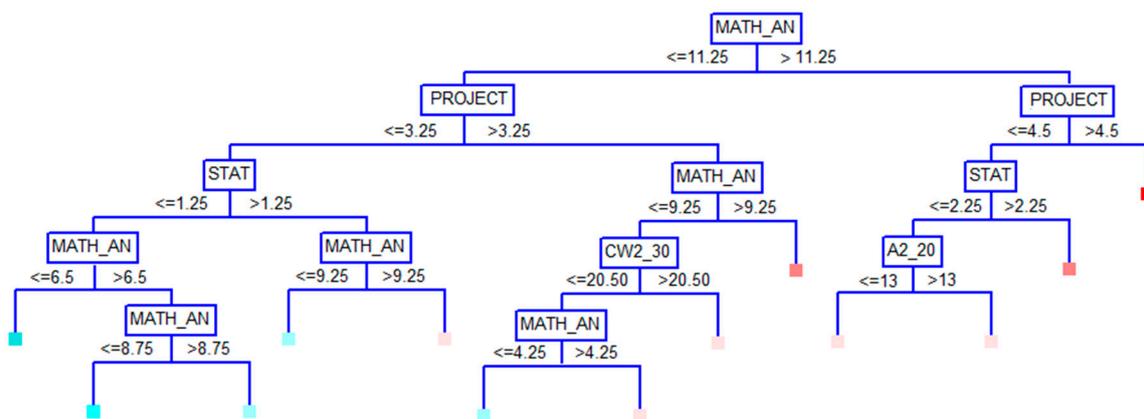

Figure 4. Diagram of the regression tree of CART model M_1 with split variables denoted and respective threshold values for each tree node.

Figure 4 presents the regression tree of model M_1 with the splitting variables and their threshold values at each split. The initial splitting variable at the tree root is *Math_An*. The rule for splitting each root case into two is "cases with a value of *Math_An* <= 11.25 go into the left child node and the rest into the right one". Here the threshold value is *Math_An* = 11.25. At the next (first) level, the splitting variable for both nodes is *Project* with the same splitting rule for the cases: "*Project* <= 4.5" etc. The process which generates the CART tree M_1, shown in Figure 4, is stopped to a depth 5, having 13 terminal nodes,

marked with a colored square. The value predicted by the model for each case of the given terminal node is the arithmetical mean of *Exam* points from the cases classified in that node.

The tree of the M_4 model shown in Figure 5 has an identical structure as the tree of the M_1 model from Figure 4. An almost complete match is observed; therefore, stability of the CART models is obtained.
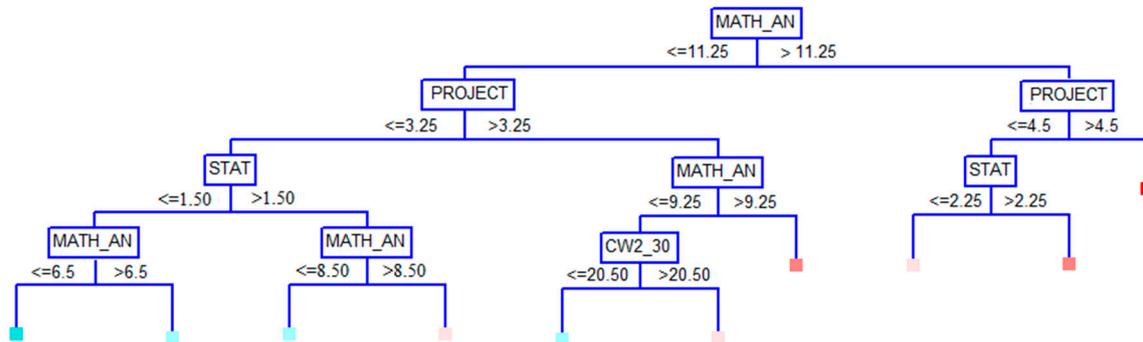

Figure 5. Diagram of the regression tree of CART model M_4 with split variables denoted and respective threshold values for each tree node.

The predicted scores obtained from the six CART models in Table 4 were statistically examined with a Wilcoxon signed rank test for paired samples. It was found that all Wilcoxon signed rank tests were statistically unsignificant and the differences of each two models had symmetrical distributions. This is an indicator that the models do not differ significantly from each other.

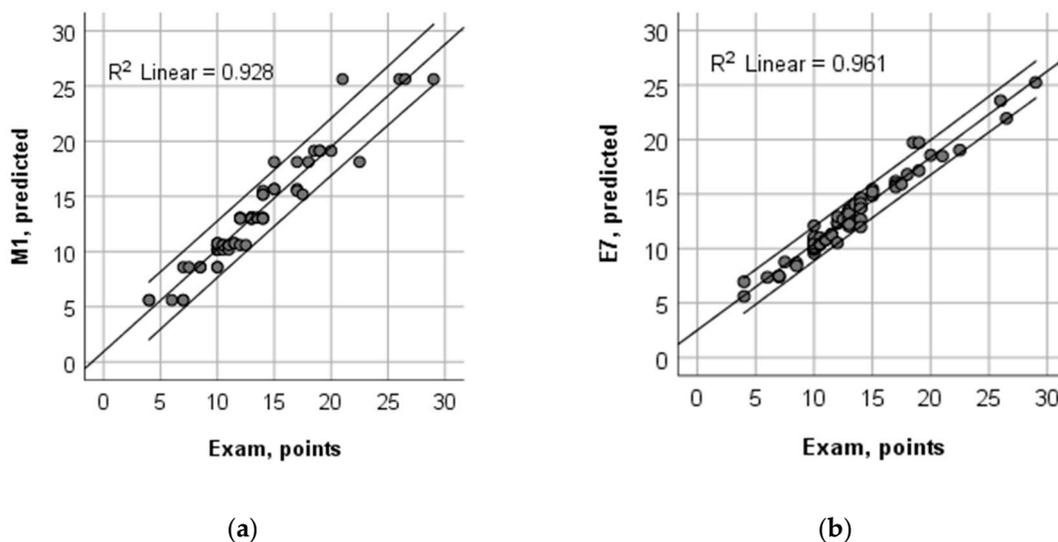Figure 6 shows the actual versus predicted values of the *Exam* obtained by model M_1.

(**a**)

(**b**)

Figure 6. Scatter plots of predicted versus actual values of the target variable *Exam* with 5% confidence interval for (**a**) CART model M_1 and (**b**) CART-EB model E_7.

Figure 7 presents line plots of *Exam* and model predictions obtained by models M_1 and E_7. For both models there are larger differences with *Exam* at the highest values.

Figure 7. Line plots of the target variable *Exam* and its predicted values with models M_1 and E_7.

3.5. Results from the CART Ensembles and Bagged Models

The same values of the hyperparameters as in the CART models were used in the construction of the CART-EB models. For the minimum number of cases in a parent node (m1) and the minimum number of cases in a child node (m2), two options were set: 5-2 and 6-3, respectively. All trees in the ensemble were trained with 10-fold cross-validation. Bagged trees were built with repeated cases. Due to the small sample size ($n = 68$), the ensembles were built with 10, 15, 20 and 25 trees. The family of these models are denoted by E_1, E_2, ..., E_8. Of these, the first group E_1, E_2 and E_3 are built using values of the hyperparameters m1-m2 equal to, respectively, 6-3. For the second group with the remaining models, these parameters are 5-2. The first models of each group, namely E_1 and E_4, are initial CART trees. Table 5 presents the statistical indicators obtained for these models. The models E_2 and E_3 show relatively low R^2 Test and the largest errors RMSE Test and Learn. Also the corresponding Theil's coefficients are larger than those of the models with m1 = 5 and m2 = 2. Model E_6 has the best test statistics with R^2 Test (0.922) and RMSE Test (1.838). The next model E_7 has the best indicators for the training sample—R^2 Learn (0.961) and RMSE Learn (1.278). As the number of terminal nodes increases, the statistics become less favorable, as seen from those of model E_8. Therefore, in further analysis we consider the models E_6 and E_7. The predictive properties of model E_7 are illustrated in Figures 6b and 7.

Table 5. Summary statistics of regression CART ensembles and bagged models for assessment of students' achievements.

Statistic	CART Ensembles and Bagged Model							
	E_1 Initial Tree	E_2	E_3	E_4 Initial Tree	E_5	E_6	E_7	E_8
Number of trees	-	10	15	-	10	15	20	25
m1-m2	6-3	6-3	6-3	5-2	5-2	5-2	5-2	5-2
R^2, Test	0.845	0.886	0.897	0.845	0.916	0.922	0.883	0.807
R^2, Learn	-	0.923	0.942	-	0.936	0.953	0.961	0.945
RMSE, Test	2.400	2.222	2.055	2.368	2.092	1.838	1.908	2.302
RMSE, Learn	-	1.724	1.583	-	1.610	1.370	1.278	1.368
Theil's U_{II}	-	0.0121	0.0132	-	0.0177	0.0104	0.0086	0.0098

The results of the Wilcoxon signed rank tests show that the built CART-EB models do not differ significantly from each other.

3.6. Combination of CART and CART Ensembles and Bagged Models Using MARS

To improve the quality of prediction, MARS regression models of the dependent variable *Exam* were built using the selected four best models M_1, M_4, E_6 and E_7 as predictors. Due to the almost linear behavior of the *Exam* curve, only a linear MARS method was applied. The MARS models generated are denoted by MM_1, MM_2 and MM_3. Furthermore, by finding the importance of individual regression trees models, it can be determined which of them has the best predictive properties. The models were trained with 10-fold cross-validation. The summary statistics of the models obtained are presented in Table 6.

Table 6. Summary statistics of regression of MARS models built using CART and CART-EB models as predictors [1].

Statistic	MARS Model		
	MM_1	MM_2	MM_3
Predictors	M_1, E_7	M_1, E_6, E_7	M_1, M_4, E_6, E_7
Number of BFs	4	5	6
Variable importance [2]	23, 100	42, 29, 100	97, 0, 55, 100
R^2 Test	0.960	0.956	0.954
R^2 Learn	0.972	0.974	0.978
GCV R^2	0.958	0.960	0.960
RMSE Test	0.966	1.021	1.035
RMSE Learn	0.804	0.749	0.726
Theil's U_{II}	0.0034	0.0029	0.0028

[1] GCV R^2 stands for general cross-validation R^2 [30,33]. [2] Variable importance corresponds to the order of the predictors in column 4.

The statistics of the built models in Table 6 are almost the same. Using all four models (see MM_3), we found that the greatest influence was exerted by model E_7 (100 relative units), followed by M_1 (97 units), E_6 (55 units) and M_4 (0 units). By successively reducing the predictors, the other models are obtained. We choose the simplest MM_1 model for optimal. MM_1 outperforms the separately taken CART and CART-EB models in all evaluation metrics from (4). The MM_1 model has the form

$$
\begin{aligned}
\hat{E}xam &= 3.64249 + 1.52178 * BF1 + 0.275286 * BF3 - 0.843117 * BF5 + 0.321376 * BF7, \\
BF1 &= max(0, E_7 - 5.605), \ BF3 = max(0, E_7 - 14.1583), \\
BF5 &= max(0, E_7 - 9.5625), \ BF7 = max(0, M_1 - 10.1875).
\end{aligned} \tag{5}
$$

Figure 8 shows the scatter plot of the actual *Exam* values versus MM_1 model predictions.
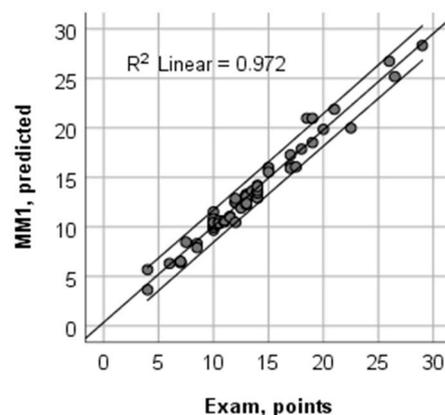

Figure 8. Scatter plot of the predicted values from optimal MARS model MM_1 versus actual values of the target variable *Exam* with a 5% confidence interval.

4. Discussion with Conclusions

This study presents, models and analyzes results from a competency-based exam in applied mathematics together with the results of preparatory academic activities. They are modeled using three ML methods—CART, CART ensembles and bagging, and MARS.

The CART method was first applied. Table 3 indicates that the six CART models selected have high goodness-of-fit indicators with coefficients of determination R^2 over 90% and RMSE around 1.5, or within 5%. As optimal models, we selected M_1 and M_4. M_1 shows $R^2 = 0.928$, RMSE = 1.298, as well as a small value of the Theil's forecast accuracy coefficient $U_{II} = 0.0089$.

The CART models allow for determining the influence of individual educational components on exam results for the specific subject of applied mathematics. The importance of the predictors in the M_1 model in relative units is *Project* (100), *Math_An* (95), *A2_20* (54), *CW1_30* (38), *A1_12* (36) etc., as presented in Table 4. Therefore, the solution of the project and the tasks of mathematical analysis determine to the greatest extent the achievements of the students. Other important factors are the grades from the second homework, the first control test etc. The influence of students' success with problem 2 in statistics (variable *Stat*) reaches only about 22% relative weight. This indicates an unsatisfactory level of theoretical knowledge in probabilities and statistics. Using these results with Table 1 of competencies, it is apparent that following a reduction of competencies for the two problems from *Stat*, the exam test can be assessed mainly by the acquisition of competencies with the most "+" and "0". These competencies are C3, C4, C6 and C8.

The data were also modeled using the CART ensembles and bagging method. Six CART-EB models were built. The analysis of these models showed that the best statistical evaluation indicators were for models E_6 and E_7, with 15 and 20 trees in the ensemble, respectively. As an optimal model, E_7 achieved $R^2 = 0.961$, RMSE = 1.278, as well as a Theil's accuracy coefficient $U_{II} = 0.0086$. Although the EB models do not derive the influence of individual predictors, they serve as confirmatory and complementary to CART models for predicting student achievement.

The idea arose to combine the four best models—two CART and two CART-EB models for regression of the dependent variable *Exam*. A linear MARS method was applied. Three models with predictors M_1, M_4, E_6 and E_7 were built. The models showed very close goodness-of-fit indices. The final model selected, MM_1, used M_1 and E_7 and achieved $R^2 = 0.972$, RMSE = 0.804 and Theil's accuracy coefficient $U_{II} = 0.0034$. This model showed a significant improvement in the prediction of the lowest and highest exam scores.

The results obtained are comparable to those obtained by us in [6], where regression-type CART models were constructed to predict the final exam results in linear algebra and analytical geometry for students in two other specialties at the same university, using a short mathematical competency test and mid-term test results. In [6], the models reached to fit the actual data with $R^2 = 93\%$. The results obtained here are also similar to those in [11,13].

It should be noted that for the first time the CART ensemble and bagging method for data from education was applied. In addition, for the first time, combining the predictions of the individual competing models using the MARS method is used. The combined MARS models obtained exceed the qualities of the predictors included in all statistical indicators.

Essentially, the approach for modeling the data we present consists of two consecutive steps: (1) building regression trees and regression trees ensemble models (Sections 3.4 and 3.5) using the initial predictors, and (2) building MARS models (Section 3.6), where the predictors are the resulting variables with values predicted in step (1). Based on this, we can formulate some advantages and potential capabilities of this approach, namely:

- As part of the family of regression trees, the CART and CART-EB methods we use can successfully deal with uncertainty, qualitatively stated problems and incomplete, imprecise or even contradictory data sets, as stated in [10]. These can process both nominal and numerical data, handle multidimensional and multivariety data, easily

identify patterns and nonlinear complex relationships between the predictors, thus facilitating the interpretation of models.

- At step (1), the variable importance of initial predictors in models is assessed directly, which allows us to ignore/screen unsignificant predictors. This would be especially useful in the case of a large number of predictors for reducing the dimensionality of the problem.

- At step (2), numerical-type data are used, enabling the implementation of the MARS method, whereby it is combined with the predictions from (1). In particular, the results of our study showed that MARS models improve the predictions of the smallest and largest values of the target variable, including its outliers. In this manner, it is possible to eliminate or reduce the effect of this type of flaw, typical for all ensemble methods.

- The importance of the models from step (1), used as predictors, is determined with the help of MARS in step (2). In this manner, the best regression trees model is identified. Indirectly, if it determines the influence of the initial predictors, additional useful information may be obtained to interpret the overall statistical analysis.

In addition to this, the proposed research method has several limitations. The models can be built if at least 50 data records are available. Furthermore, the CART-EB algorithm used does not deduce the relative importance of the predictors in the model, which makes any direct interpretation of the results difficult. Another disadvantage typical for all ML methods is that results depend to a certain degree on accuracy criterion, variable and model selection.

The proposed methods and models in this study can be used to direct and improve exam tests for students in subsequent courses, making changes at the tutor's discretion. Changes can be made both in the educational content, tests and the academic programs, and the management of other basic factors that influence grades, as determined using the models. This approach promises to find hidden relationships between factors contributing to learning and teaching, and also benefits tutors/authorities by making predictions and helping them make better decisions. Future research can be planned in this regard. By applying the approach we propose, another crucial practical issue for further research is determining the factors and predicting which students may drop out.

We can conclude that the use of small practical projects as a competency-oriented approach and combined with the application of powerful ML methods for processing the data set related to the learning process are effective for assessment of students' knowledge and competencies in mathematics.

Author Contributions: Conceptualization and methodology, all authors; data preparation, H.K. and A.I.; modeling, S.G.-I.; validation, H.K. and A.I.; analysis of the results, all authors; writing—review and editing, S.G.-I. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abdulwahed, M.; Jaworski, B.; Crawford, A. Innovative approaches to teaching mathematics in higher education: A review and critique. *Nord. Stud. Math. Educ.* **2012**, *17*, 49–68. Available online: https://repository.lboro.ac.uk/articles/Innovative_approaches_to_teaching_mathematics_in_higher_education_a_review_and_critique/9370940/files/16981556.pdf (accessed on 17 November 2020).
2. Hassan, O.A.B. Learning theories and assessment methodologies—An engineering educational perspective. *Eur. J. Eng. Educ.* **2011**, *36*, 327–339. [CrossRef]
3. Niss, M. Mathematical Competencies and the Learning of Mathematics: The Danish KOM Project. In Proceedings of the 3rd Mediterranean Conference on Mathematical Education, Athens, Greece, 3–5 January 2003; The Hellenic Mathematical Society: Athens, Greece, 2003; pp. 115–124. Available online: http://www.math.chalmers.se/Math/Grundutb/CTH/mve375/1112/docs/KOMkompetenser.pdf (accessed on 17 November 2020).

4. Alpers, B.A.; Demlova, M.; Fant, C.-H.; Gustafsson, T.; Lawson, D.; Leslie Mustoe, L.; Olsen-Lehtonen, B.; Robinson, C.; Velichova, D. *A Framework for Mathematics Curricula in Engineering Education: A Report of the Mathematics Working Group*; European Society for Engineering Education (SEFI): Brussels, Belgium, 2013; Available online: http://sefi.htw-aalen.de/Curriculum/Competency%20based%20curriculum%20incl%20ads.pdf (accessed on 17 November 2020).

5. Queiruga-Dios, A.; Hernández Encinas, A.; Demlova, M.; Dias Rasteiro, D.; Rodríguez Sánchez, G.; Sánchez Santos, M.J. Rules_Math: Establishing Assessment Standards. In *Advances in Intelligent Systems and Computing*; Martínez Álvarez, F., Troncoso, L.A., Sáez Muñoz, J., Quintián, H., Corchado, E., Eds.; Springer: Cham, Switzerland, 2020; Volume 951, pp. 235–244. [CrossRef]

6. Gocheva-Ilieva, S.; Teofilova, M.; Iliev, A.; Kulina, H.; Voynikova, D.; Ivanov, A.; Atanasova, P. Data Mining for Statistical Evaluation of Summative and Competency-Based Assessments in Mathematics. In *Advances in Intelligent Systems and Computing*; Martínez Álvarez, F., Troncoso, L.A., Sáez Muñoz, J., Quintián, H., Corchado, E., Eds.; Springer: Cham, Switzerland, 2020; Volume 951, pp. 207–216. [CrossRef]

7. Ivanov, A. Decision trees for evaluation of mathematical competencies in the higher education: A case study. *Mathematics* **2020**, *8*, 748. [CrossRef]

8. Neumann, I.; Rosken-Winter, B.; Lehmann, M. Measuring mathematical competences of engineering students at the beginning of their studies. *Peabody J. Educ.* **2015**, *90*, 465–476. [CrossRef]

9. Georgieva, P.V.; Nikolova, E.P. Enhancing communication competences through mathematics in engineering curriculum. In Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019, Opatija, Croatia, 20–24 May 2019; Volume 8757207, pp. 1451–1456. [CrossRef]

10. Charitopoulos, A.; Rangoussi, M.; Koulouriotis, D. On the use of soft computing methods in educational data mining and learning analytics research: A review of years 2010–2018. *Int. J. Artif. Intell. Educ.* **2020**, *30*. [CrossRef]

11. Mesarić, J.; Šebalj, D. Decision trees for predicting the academic success of students. *Croat. Oper. Res. Rev.* **2016**, *7*, 367–388. [CrossRef]

12. Kotsiantis, S.; Pierrakeas, C.; Pintelas, P. Predicting students' performance in distance learning using machine learning techniques. *Appl. Artif. Intell.* **2004**, *18*, 411–426. [CrossRef]

13. Mueen, A.; Zafar, B.; Manzoor, U. Modeling and predicting students' academic performance using data mining techniques. *Int. J. Mod. Educ. Comp. Sci.* **2016**, *8*, 36–42. [CrossRef]

14. Behr, A.; Giese, M.; Teguim, K.; Theune, K. Early prediction of university dropouts—A random forest approach. *J. Econ. Stat.* **2020**, *240*, 743–789. [CrossRef]

15. Sokkhey, P.; Okazaki, T. Comparative study of prediction models for high school student performance in mathematics. *IEIE Trans. Smart Process. Comput.* **2019**, *8*, 394–404. [CrossRef]

16. Sokkhey, P.; Okazaki, T. Hybrid machine learning algorithms for predicting academic performance. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 32–41. [CrossRef]

17. Qiang, T. Data mining algorithm and the effectiveness of mathematics classroom teaching based on support vector machine. *Int. J. Database Theory Appl.* **2016**, *9*, 163–174. [CrossRef]

18. Siri, A. Predicting students' dropout at university using artificial neural networks. *Ital. J. Soc. Educ.* **2015**, *7*, 225–247. [CrossRef]

19. Mat, U.B.; Buniyamin, N. Using neuro-fuzzy technique to classify and predict electrical engineering students' achievement upon graduation based on mathematics competency. *Indones. J. Electr. Eng. Comput. Sci.* **2017**, *5*, 684–690.

20. Ivanova, V.; Zlatanov, B. Implementation of fuzzy functions aimed at fairer grading of students' tests. *Educ. Sci.* **2019**, *9*, 214. [CrossRef]

21. Depren, S.K. Prediction of students' science achievement: An application of multivariate adaptive regression splines and regression trees. *J. Balt. Sci. Educ.* **2018**, *17*, 887–903. [CrossRef]

22. Shahiri, A.M.; Husain, W.; Rashid, N.A. A review on predicting student's performance using data mining techniques. *Proced. Comp. Sci.* **2015**, *72*, 414–422. [CrossRef]

23. Dutt, A.; Ismail, M.A.; Herawan, T. A systematic review on educational data mining. *IEEE Access* **2017**, *5*, 15991–16005. [CrossRef]

24. Alyahyan, E.; Düştegör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]

25. Pintelas, P.; Livieris, I.E. Special issue on ensemble learning and applications. *Algorithms* **2020**, *13*, 140. [CrossRef]

26. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984.

27. Steinberg, D. CART: Classification and regression trees. In *The Top Ten Algorithms in Data Mining*; Wu, X., Kumar, V., Eds.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2009; pp. 179–202.

28. Izenman, A.J. *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*; Springer: New York, NY, USA, 2008.

29. Apté, C.; Weiss, S. Data mining with decision trees and decision rules. *Future Gener. Comput. Syst.* **1997**, *13*, 197–210. [CrossRef]

30. Salford Predictive Modeler. Available online: https://www.minitab.com/en-us/products/spm/ (accessed on 17 November 2020).

31. Wen, S.; Buyukada, M.; Evrendilek, F.; Liu, J. Uncertainty and sensitivity analyses of co-combustion/pyrolysis of textile dyeing sludge and incense sticks: Regression and machine-learning models. *Renew. Energy* **2019**, *151*, 463–474. [CrossRef]

32. Pradeepkumar, D.; Ravi, V. Forex rate prediction: A hybrid approach using chaos theory and multivariate adaptive regression splines. In *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 515, pp. 219–227. [CrossRef]
33. Friedman, J.H. Multivariate adaptive regression splines (with discussion). *Ann. Stat.* **1991**, *19*, 1–141. [CrossRef]
34. Bliemel, F. Theil's forecast accuracy coefficient: A clarification. *J. Mark. Res.* **1973**, *10*, 444–446. [CrossRef]
35. Vandekerckhove, J.; Matzke, D.; Wagenmakers, E.J. Model comparison and the principle of parsimony. In *The Oxford Handbook of Computational and Mathematical Psychology*; Busemeyer, J.R., Wang, Z., Townsend, J.T., Eidelsm, A., Eds.; Oxford University Press: Oxford, UK, 2015; pp. 300–318.
36. Queiruga-Dios, A.; Sanchez, M.J.S.; Perez, J.J.B.; Martin-Vaquero, J.; Encinas, A.H.; Gocheva-Ilieva, S.; Demlova, M.; Rasteiro, D.D.; Caridade, C.; Gayoso-Martinez, V. Evaluating Engineering Competencies: A New Paradigm. In Proceedings of the Global Engineering Education Conference (EDUCON), Tenerife, Spain, 17–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 2052–2055. [CrossRef]