Sigma J Eng & Nat Sci 38 (2), 2020, 963-977

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article

EDUCATIONAL DATA MINING METHODS FOR TIMSS 2015 MATHEMATICS SUCCESS: TURKEY CASE

Enes FİLİZ[1], Ersoy ÖZ*[2]

[1]*Fırat University, Department of Statistics, ELAZIĞ;* ORCID: 0000-0002-8006-9467
[2]*Yildiz Technical University, Department of Statistics, İSTANBUL;* ORCID: 0000-0001-9087-434X

Received: 17.01.2020 Revised: 06.05.2020 Accepted: 07.05.2020

ABSTRACT

Educational data mining (EDM) is an important research area which has an ability of analyzing and modeling educational data. Obtained outputs from EDM help researchers and education planners understand and revise the systematic problems of current educational strategies. This study deals with an important international study, namely Trends International Mathematics and Science Study (TIMSS). EDM methods are applied to last released TIMSS 2015 8th grade Turkish students' data. The study has mainly twofold: to find best performer algorithm(s) for classifying students' mathematic success and to extract important features on success. The most appropriate algorithm is found as logistic regression and also support vector machines - polynomial kernel and support vector machines - Pearson VII function-based universal kernel give similar performances with logistic regression. Different feature selection methods are used in order to extract the most effective features in classification among all features in the original dataset. "Home Educational Resources", "Student Confident in Mathematics" and "Mathematics Achievement Too Low for Estimation" are found the most important features in all feature selection methods.
Keywords: Classification algorithms, educational data mining, feature selection, mathematics success, Timss 2015.

1. INTRODUCTION

Educational Data Mining (EDM) offers educators and educational planners a better understanding of big and complex educational datasets. With basic statistical analysis, one can interpret the general behaviors of datasets. However, discovering the useful information which is unknown or inaccessible can be done only by using EDM methods. The extracted useful information can be used for educators, educational planners and decision makers in the field of education. Also, these information monitor the current situation of student's successes, and thus solutions can be offered to the defective educational strategies.

In order to use EDM with effectively, studying with reliable datasets plays a vital role. The International Association for the Evaluation of Educational Achievement (IEA) carries out many comparative studies to obtain and report the reliable datasets with the aim of monitoring the effects of policies and practices of education systems for participating countries. Trends in International Mathematics and Science Study (TIMSS) is one of the biggest project of IEA that

* Corresponding Author: e-mail: ersoyoz@yildiz.edu.tr, tel: (212) 383 72 65

assess the student success at international level and more than 60 countries have participated to this project. The project is realized every four years with the fourth and eighth-grade mathematics and science students. TIMSS not only provides the reliable information regarding the effects of education policies and their implementations but also it enables researchers to make comparisons among the participating countries' results in terms of student success [1].

Although important studies have been done with using TIMSS data, most of them used different classic statistical methods such as regression, factor analysis and etc. However, these methods have some limitations, especially for the big and complex datasets. Besides, the absence of classical statistical assumptions such as normality, variance homogeneity and linearity, has made EDM popular [2, 3]. Thus, this study aims to contribute the current literature of TIMSS studies which are in the concept of EDM. For this purpose, mainly two research questions are being fully addressed: (1) which EDM method gives the best performance to classify the TIMSS data (2) what are the important features (factors) related to students' success.

The first research question is more likely to help guide researchers in determining the most appropriate algorithms in classification within the concept of EDM. EDM has various methods to model, analyze and interpret the datasets. In this study, k-nearest neighbor (k-NN), naïve Bayes (NB) algorithm, C4.5 decision tree (DT-C4.5) algorithm, raptree decision tree (DT-RepTree) algorithm, random forest decision tree (DT-RF) algorithm, artificial neural networks (ANN), support vector machines (SVM) with three different kernels which are polynomial kernel (SVM-POLY), radial basis function kernel (SVM-RBF) and Pearson VII function-based universal kernel (SVM-PUK), and logistic regression (LR) are used. These algorithms are the most used methods in EDM literature.

The second research question is about finding the most important features in classification. Since education is affected by various features, the application of EDM algorithms for student performances is essential to recognize the poor performance of students and to investigate the impacts of effective features on education [4]. The original TIMSS dataset includes many potential affecting features to students' success and it is very important to put forward the most influenced ones. The correct identification helps decision makers to adjust the current defective situation. In this work, different feature selection algorithms such as correlation based, correlation analysis, scale based and sensitivity analysis are applied in order to extract effective features on students' success. Also, it is essential to keep the number of variables in the classification algorithms as low as possible while maintaining the maximum classification performance.

2. LITERATURE REVIEW

EDM is an important research area in analyzing, modeling and future decision making based on educational data. EDM is used to understand the datasets, students and their learning process better. It is also used for developing practical approaches to provide useful information for the students. EDM has been drawing attention as a research area for researchers all over the world in recent years [5]. Romero and Ventura [6] studied educational data by conducting a comprehensive literature search. In their work, they showed the EDM as a recursive loop involving hypothesis building, testing and performance improvement. Also, the study mentioned that the results of the EDM methods can lead the educators to discover useful information about evaluations. In other important study of Romero and Ventura [5], they included general definitions of EDM. Then, they sorted out common tasks in the educational environment that were solved by methods of data mining. They presented discussions about the future possibilities according to the results. Baker and Yacef [7] investigated the historical and current changes in EDM field. They examined the studies conducted by the society on the basis of the methodological definition of research in the first years of EDM. Siemens and Baker [8] examined the simultaneous application of data mining and analytical methods for the development of two research communities, EDM and learning analytics and knowledge. Mohamad and Tasir [9]

aimed at exam ining the latest trends in data mining in educational research and how data mining was addressed by previous researchers. Some points of the current research were discussed, and new ideas for future research were proposed. Peña-Ayala [10] followed a two-stage road in his study. The first was to maintain and promote the history of EDM's development. The second was to analyze and discuss the results, depending on the conclusions reached by a data mining approach. In this way, the selection and analysis of 240 EDM works were made.

EDM methods have many different algorithms in order to classification, estimation and prediction of datasets. Shahiri and Husain [11] systematically searched the literature to find performance estimation methods for the most successful students from 2002 to 2015. They showed that the most commonly used EDM methods were ANN, DT, SVM, k-NN and NB respectively. Kotsiantis, Pierrakeas and Pintelas [12] compared some advanced EDM methods. It was determined that NB algorithm was the most suitable method to use in the construction of a software support tool by analyzing the results. Ramaswami and Bhaskaran [4] conducted a survey to investigate the academic performance and applied it to students and school principals. They obtained some prediction rules using the chi-square automatic interaction detection model. It was shown that this prediction model is effective and gave good results compared to the other models. Baradwaj and Pal [13] aimed to provide a data mining model for the higher education system, demonstrating that data mining algorithms are adequate. In this study, the classification method was used to evaluate the performance of the student. Besides, since there were many approaches used in data classification, the decision tree method was selected for the application. Rajni and Malaya [14] presented ideas about the utility and availability of EDM to solve the engineering educational planning problem. They chose decision trees and ANN methods as prediction models. In the study, they obtained forecasting models for Indian engineering entrance examination data. Martínez Abad and Chaparro Caso López [15] presented a procedure used to elicit academic success factors in the direction of the emergence of statistical analysis techniques based on data mining in educational sciences. It was demonstrated that personal factors were the most indicative of academic performance based on the results obtained using DT. As a result, they emphasized the importance of DT allowing a better conclusion and interpretation than other models and techniques. Cortez and Silva [16] used four machine learning methods, DT, RF, ANN and SVM, to build a model based on student performance in secondary schools in Portugal. Findings not only showed the best prediction model but also supported the idea that academic success and past performance were highly correlated with each other. Osmanbegović and Suljić [17] used data mining methods to predict the success levels of Tuzla University students. Web-based learning system was proven to be an important part of the development of the student's creativity.

Most of studies that mentioned above were designed for at national education data. Educational data which includes the assessment of many countries' results in education such as TIMSS helps researchers to compare/improve educational policies and their results in terms of success. Several important studies dealt with modeling the TIMSS data by EDM methods. Hammouri [18] aimed to find the factors influencing students' mathematical success. According to the results obtained with the help of the TIMSS dataset, attitude, achievement, confidence in mathematical ability and perception of the significance of mathematics determine the student's mathematics success. Liu and Meng [19] studied the factors in TIMSS 2003 dataset and examined the mathematical awareness concept of high and low achieving students in East Asia and America using these factors. Askin and Gokalp [20] examined the factors that effect the educational success of students for TIMSS 2011 data. They used LR and ANN methods to measure prediction and classification performance. The most effective factor was determined as students' confidence. Topçu, Erbilgin and Arıkan [21] used TIMSS 2011 data to investigate factors affect the Turkish and Korean students' success in science and mathematics. Also, they discussed the educational implications according to their findings. Kılıç-Depren, Askin and Öz [22] applied DT, NB, LR and ANN to TIMSS 2011 data. They aimed to find the best performing algorithm for classification of eighth-grade Turkish students using various performance measures for their

mathematical success. Filiz and Oz [23] applied EDM methodology on TIMSS 2015 science data. They found that which factors are most effective in science success.

3. MATERIAL AND METHODS

3.1. Data set

TIMSS has been conducted for every 4 years to assess mathematics and science achievement and the target populations are 4th and 8th grade students. In TIMSS 2015, 39 participating countries and 7 benchmarking participants attended to 8th grade assessment. TIMSS sampling procedure is a two-stage random sample design. The first stage is selecting a sample of schools proportionally and the second is randomly selecting classes among sampled schools. This project is a survey based on self-reported questionnaires [24].

Table 1. Student Related Features

Factor Name	Description	Factor Name	Description
ITSEX	Sex of students	*BSBG13C*	How often use computer tablet\other
BSBG03	Often speak at home	*BSBG14A*	Access textbooks
BSBG05	Digital information devices	*BSBG14B*	Access assignments
BSBG06A	Computer tablet own	*BSBG14C*	Collaborate with classmates
BSBG06B	Computer tablet shared	*BSBG14D*	Communicate with teacher
BSBG06C	Study desk	*BSBG14E*	Find info to aid in math
BSBG06D	Own room	*BSBM38AA*	Find info to aid in math
BSBG06E	Internet connection	*BSBM39AA*	Extra lessons last 12 month\mathematics
BSBG06F	Own mobile phone	*BSBM39BA*	Extra lessons how many month\mathematics
BSBG06G	Gaming system	*BSBGHER*	Home educational resources
BSBG06H	Heating systems	*BSBGSSB*	Students sense of school belonging
BSBG06I	Cooling systems	*BSBGSB*	Student bullying
BSBG06J	Washing machine	*BSBGSLM*	Students like learning mathematics
BSBG06K	Dishwasher	*BSBGEML*	Engaging teaching in math lessons
BSBG08	How far in education do you expect to go	*BSBGSCM*	Student confident in mathematics
BSBG11	About how often absent from school	*BSBGSVM*	Students value mathematics
BSBG12	How often breakfast on school days	*BSDMLOWP*	Mathematics achievement too low for estimation
BSBG13A	How often use computer tablet\home	*BSDMWKHW*	Weekly time spent on math homework
BSBG13B	How often use computer tablet\school	*BSMMAT01*	1st plausible value mathematics

The data set of the current study is taken by TIMSS 2015 assessment for 8th grade Turkish students' mathematics results. The average mathematics score of 8th grade Turkish students is 458. Data set includes 6079 students' information for 2943 females and 3136 males. However, there are some missing and inaccurate values so the total of 4577 students' information for 2303 females and 2274 males is taken into account. All features are used in modeling the data set. However, some variables such as "age" is removed from the dataset due to lack of variability. Thus, 36 features are used and given with their descriptions in Table 1. From these variables, "1st

Plausible Value Mathematics" (BSMMAT01) feature is the students' mathematics success and it is chosen as dependent variable. The international TIMSS 2015 mathematics score has the average value of 500 with the standard deviation 100 [1]. Therefore, average value of TIMSS study is used as the center point and encoded the students' "1st Plausible Value Mathematics" score as 1 if it is higher than 500 and as 0 otherwise. Other 35 variables were defined as independent variables, in other words these variables are taken as potential important features in students' mathematics success.

In TIMSS 2015, the questionnaire was developed to combine to determine a single hidden structure. This structure is called scale. Rasch partial credit model which is one of the Item Response Theory (IRT) scaling methods was used in the reporting [25]. In this study, there are 9 scales and they consist of "Home Educational Resource" (BSBGHER), "Students Sense of School Belonging" (BSBGSSB), "Student Bullying" (BSBGSB), "Students Like Learning Mathematics" (BSBGSLM), "Engaging Teaching in Math Lessons" (BSBGEML), "Student Confident in Mathematics" (BSBGSCM), "Students Value Mathematics" (BSBGSVM), "Mathematics Achievement Too Low for Estimation" (BSDMLOWP) and "Weekly Time Spent on Math Homework" (BSDMWKHW) variables.

3.2. k-fold Cross-validation

One of the critical points when applying data mining methods is k-fold Cross-validation. The dataset is mainly divided into two groups as training and testing set. The division can be determined by partitions such as 50% - 50%, 70% - 30%. In addition, the k-fold Cross-validation can be used to divide the data set into k pieces; k-1 of them for training, and the remaining one for testing. This process is repeated k times by taking each part as a testing set. When the mean of all result is calculated, the values of classification measures are obtained [26].

3.3. Classification Algorithms

k-Nearest neighbor algorithm (k-NN): The k-NN algorithm is a sample-based learning method that uses all training data for classification [27]. To perform classification in a data set, it finds the closest neighbors among the variables. The distances between the data points are important. Moreover, the performance of this algorithm depends on the variables, that is, the value for an appropriate similarity function and the parameter k [28].

Naive bayes algorithm (NB): The NB classification algorithm is a particular form of Bayesian networks, and two assumptions must be satisfied. The first assumption is, the classes must be conditionally independent. The second assumption is that the variables that affect the final result are not hidden [29]. The NB classifier is distinguished as one of the most efficient inductive learning algorithms for data mining [30].

Support vector machines (SVM): SVM generate an n-dimensional hyperplane that optimally divides the data into two categories [31]. If the data is linearly separated, linear SVM is used; if it cannot be linearly separated, then the non-linear SVM is used [32]. The primary task of the nonlinear SVM approach lies in the selection of the kernel function. Linear, radial basis, polynomial, sigmoid, second-order multiple, reverse second-order kernel functions are used in general. Choosing different kernel functions produce different SVM and can lead to different performance results [33]. The correct choice of kernel function has a substantial influence on learning capacity [34].

Artificial neural networks (ANN): ANN mimics the processes of the human brain. ANN models can be learned and generalized according to the results obtained from past experiences, trained like a human brain. The most important advantage of using ANN is that even when working with complex nonlinear relationships, they do not need strict assumptions, as in standard statistical

methods. ANN use the backpropagation algorithm in the training process. The model has mainly three components: the input layer, hidden layers and an output layer [2].

Decision trees (DT): The main purpose of using DT algorithms is to construct a decision tree out of a given dataset by minimizing the generalization error [35]. DT algorithms are one of the important classification techniques due to their interpretable character. The most used DT algorithms are C4.5, RepTree and RF. DT-C4.5 is known as J48 in an open source Weka application of C4.5. It produces a binary decision tree using information entropy. This method can be successfully used in many pattern recognition problems [36]. DT-RepTree approach is designed by regression tree logic. The algorithm makes various trees in different iterations and selects the best tree from all the trees made. The mean square error criterion is used for tree trimming and selection [37]. DT-RepTree algorithm is a quick decision tree learning. It also creates a decision tree based on knowledge acquisition or reduction of variance. It only takes the values of the numerical attributes once and uses fractional samples of C4.5 for missing observations [38]. In DT-RF method, many decision trees are created using different variations of a training data. New versions of the training data are obtained by arbitrarily replacing from the original training data set and selecting an example. Every tree in the forest should be advanced to the deepest possible level without pruning. In order to classify a new test substance, each tree in the forest is allowed to decide on classification. As a result, a classification decision is made for the majority of the cases [39]. It is also a better method in terms of performance than the equivalent algorithms [40].

Logistic regression (LR): As in standard regression models, the relationships between dependent and independent variables can be investigated using LR. The most critical assumption in the standard regression is that the dependent variable must be continuous. If the dependent variable has a value of 0 or 1, binary LR can be applied to observable independent variables to estimate or classify the dependent variable [41].

3.4. Classification Criteria

In the process of determining the superior algorithms, the results of different classification criteria are calculated and compared. In this study, true positive (TP) rate, false positive (FP) rate, precision, F-measure, Kappa (κ) statistic, mean absolute error (MAE) and root mean square error (RMSE) are used as classification criteria.

TP rate is obtained by the ratio of correctly classified positive samples to the total number of positive samples in the model. FP rate is determined by the ratio of the total number of negative samples which are actually negative but classified as positive to the total number of negative samples. The precision value is found by the ratio of the correctly classified positive samples to the total number of positive prediction samples. The F-measure value is determined by the harmonic mean of the precision value and the TP rate value [42]. κ statistic quantifies the predictive performance of a classification model. It is an appropriate statistic to measure the agreement for categorical variables. It is also a value based on the chi-square table [43]. p_o and p_e show the relation between two categorical variables [44]. MAE and RMSE statistics help to demonstrate the differences between the predicted and observed values of a model [45]. The MAE calculates the mean of the absolute differences between the predicted and observed values. RMSE is equal to the mean square root of the squared differences between the estimated and observed values. P_i and O_i show the predicted and observed values respectively. $P_i - O_i$ represents the prediction error of the model [22].

$$TPRate = \frac{TP}{TP+FN} \qquad FPRate = \frac{FP}{FP+TN} \qquad Precision = \frac{TP}{TP+FP} \quad \kappa = \frac{p_o - p_e}{1 - p_e}$$

$$F - measure = \frac{2 * Precision * TPRate}{Precision + TPRate} \quad MAE = n^{-1} \sum_{i=1}^{n} \left| P_i - O_i \right| \qquad RMSE = \sqrt{n^{-1} \sum_{i=1}^{n} \left| P_i - O_i \right|^2}$$

3.5. Feature Selection

One of the most critical points when classifying algorithms is to investigate which feature is effective. It is extremely important that the number of features used in the classification algorithms is reduced as much as possible (as in parsimony principle) and there is no significant decrease in the classification results when doing so. In this study, most important features are extracted by using different feature selection methods in order to find common features.

Correlation-based: This algorithm aims to find the best feature set by evaluating the feature sets with correlation. It tries to select a set of features with low correlation between them and features with high correlation with class tags [46, 47].

ReliefF: The ReliefF feature selection method tries to determine the values of the features and whether there are dependencies between them. To achieve this result, it compares the closest samples in the classes which the sample belongs and does not belong with weighting. It was originally developed for binary-class problems and adapted to multi-class problems [48].

Info gain: The Info gain feature selection algorithm is one of the methods that determine which attribute to start from the first branch in decision trees and is often used in filter modeled feature selection processes [47]. This method is related to concept of entropy. Entropy can be expressed as a measure of disorder in a system. The classification process reveals how much information can be gained by using the features used. The more independent the attribute from the classes is, the lower the information gain [49, 50].

One-R: Set of classification rules over the tested features is derived by One R feature selection algorithm based on a single feature value. In the One R algorithm, the key point is to select the feature with the lowest error rate. Consequently, the proportion of samples that do not belong to the majority of the feature value will contribute to the error rate [51]. The One R algorithm tests the whole data to form decision trees with certain rules. It is usually a convenient method of studying the structure of the data, which results in high accuracy [52].

Sensitivity analysis: Sensitivity analysis calculates the rate of change of model output as a result of changing the input values. The resulting estimates are used to determine the significance of each input variables [53, 54]. Sensitivity analysis can be done with many methods. One of these is sensitivity analysis with the help of layers used in ANN. It is calculated by the following formula:

$$I_i = \sum_k \frac{{w^1}_{ik}}{max_{alli,k}({w^1}_{ik})}$$

where I_i is the sensitivity of the ith node, ${w^1}_{ik}$ is the connection weight of the ith node in the first layer and kth node in the hidden layer [55]. Sensitivity analysis method helps to predict the effect of change in inputs to the change of outputs. This analysis is applied in order to determine the more important and sensitive parameters to get the accurate and precise output [53, 56].

Correlation analysis: The correlation analysis is a basic statistical technique that shows the relationship between independent variables. To show relationship between the features, correlation coefficients are calculated. Highly correlated features (with high correlation

coefficient) are excluded from the data set because uncorrelated features produce better classification [57].

3.6. Experimental Setup and Application

The experimental set up of this study can be given as follows:

- TIMSS 2015 results of Turkish 8th grade students are used as the focus group. The data set is cleaned from the not-available and missing observations.
- Training and testing data sets are obtained by using 10-fold cross validation.
- The best performing algorithm is found with using all potential influencing features. This process is explained in Phase 1.
- The most important features in success are extracted by using 6 different feature selection algorithms and the best classifiers are found by using these effective features.

In addition to feature selection algorithms, scales are used as important features. These processes are explained with Phase 2 to Phase 5. In order to investigate the research questions of this study, Weka [58] which is a Java-based and open source software developed by University of Waikato for implementations of data mining algorithms is used.

The classification of 8th grade mathematics students' success is performed using the dataset with k-NN, NB, SVM-POLY, SVM-RBF, SVM-PUK, ANN, DT-C4.5, DT-RepTree, DT-RF and LR algorithms. The application consists of 5 phases as defined below:

Phase 1: Performances of algorithms based on classification criteria is obtained by using all 35 variables and given in Table 2. The purpose of performing this phase is to show all algorithms' performances and to compare with all other phases (Phase 2 to Phase 5) in order to show the importance of feature selection.

Phase 2: In this phase different feature selection methods which are correlation-based, ReliefF, info gain and one R algorithms are applied to all 35 features. From selection methods, the highest accurate classification success is obtained by info gain algorithm. Thus, most effective features which are extracted by using info gain are reported and the analyses are re-performed. With using new feature set, the values of classification criteria for all classification algorithms are obtained.

Phase 3: Standard correlation analysis is used to find and report the important features in classification of students' success. The main idea of performing correlation analysis is to exclude highly correlated features from the dataset. Thus, the use of similar features in classification is prevented. After the most important features are extracted, classification algorithms are applied and their performances based on classification criteria are obtained.

Phase 4: Scales that are explained in dataset section is used as the most important features in this phase. The classification is achieved with scales and the algorithms' classification performances are found.

Phase 5: In this phase, sensitivity analysis is used in order to find and report the most important features. As in other phases, the values of classification criteria for all classification algorithms are obtained with using related features.

4. RESULTS

Classification performances of algorithms based on classification criteria such as TP Rate, FP Rate, Precision, F-measure, κ statistic, MAE and RMSE are given in Table 2 to Table 6. Finally, the summarized results were shown in Table 7. In order to determine the superior algorithm in each phase, first of all, the algorithm which gives the best performance based on classification criteria is selected and pointed in bold. After, the most selected algorithm is chosen as the superior algorithm of that phase.

Results of Phase 1 are given in Table 2. As it is seen, LR is the most selected algorithm according to FP Rate (0,231), F-Measure (0,802), κ statistic (0,581) and RMSE (0,368). Also, DT-RF and SVM-POLY have similar performances when compared with LR.

Table 2. Results of phase 1 with 35 features

Classifier	TP Rate	FP Rate	Precision	F-Measure	κ statistic	MAE	RMSE
k-NN	0,626	0,397	0,632	0,628	0.226	0.374	0.611
NB	0,753	0,241	0,765	0,755	0.497	0.261	0.432
DT-C4.5	0,749	0,278	0,749	0,749	0.472	0.278	0.471
DT-RepTree	0,768	0,278	0,766	0,765	0.502	0.300	0.413
DT-RF	**0,805**	0,242	**0,804**	0,801	0.579	0.303	0.376
ANN	0,753	0,275	0,752	0,753	0.480	0.250	0.479
SVM-POLY	0,804	0,239	0,803	0,801	0.579	**0.196**	0.442
SVM-RBF	0,790	0,262	0,789	0,786	0.545	0.210	0.458
SVM-PUK	0,780	0,283	0,780	0,773	0.518	0.222	0.469
LR	0,803	**0,231**	0,802	**0,802**	**0.581**	0.270	**0.368**

In Phase 2, according to Info gain feature selection method, 4 most effective features are extracted as BSBG08, BSBGHER, BSBGSCM and BSDMLOWP, respectively. With using these features, the performances of classification algorithms are obtained and given in Table 3. Based on classification criteria, DT-C4.5 and SVM-POLY are determined as the superior algorithms. Besides, classification criteria of LR have the values that are very close to DT-C4.5 and SVM-POLY.

Table 3. Results of phase 2 with 4 features

Classifier	TP Rate	FP Rate	Precision	F-Measure	κ statistic	MAE	RMSE
k-NN	0,762	0,291	0,759	0,757	0.485	0.292	0.430
NB	0,785	0,251	0,783	0,783	0.542	0.291	0.391
DT-C4.5	0,789	**0,248**	0,787	**0,787**	**0.549**	0.303	0.393
DT-RepTree	0,783	0,258	0,781	0,781	0.536	0.299	0.393
DT-RF	0,765	0,275	0,763	0,763	0.498	0.292	0.415
ANN	0,788	0,256	0,786	0,785	0.544	0.297	0.388
SVM-POLY	**0,791**	0,265	**0,791**	0,786	0.545	**0.208**	0.457
SVM-RBF	0,769	0,327	0,786	0,753	0.477	0.230	0.480
SVM-PUK	0,789	0,267	0,789	0,784	0.540	0.211	0.459
LR	0,790	0,256	0,789	**0,787**	0.548	0.300	**0.386**

In Phase 3, the correlation analysis are applied as a feature selection method and as a result of the analysis, BSBG06F, BSBG06K, BSBG08, BSBG13A, BSBG14B, BSBG14C, BSBGHER, BSBGSSB, BSBGSB, BSBGSLM, BSBGEML, BSBGSCM, BSBGSVM, BSDMLOWP and BSDMWKHW are found to be the most effective features, respectively. When the analyses are re-performed with using these 15 features, the obtained performances of classification algorithms are given in Table 4. LR is the most selected algorithm for all selection criteria except MAE. Also, SVM-POLY has similar performance compared with LR.

In Phase 4, the most important features are obtained by using scales. These scales are BSBGHER, BSBGSSB, BSBGSB, BSBGSLM, BSBGEML, BSBGSCM, BSBGSVM, BSDMLOWP and BSDMWKHW. The performances of classification algorithms which are obtained by using these 9 features are given in Table 5. As it is seen, SVM-PUK is the most

selected algorithm. Also, LR and SVM-POLY have similar performances when compared with SVM-PUK.

Table 4. Results of phase 3 with 15 features

Classifier	TP Rate	FP Rate	Precision	F-Measure	κ statistic	MAE	RMSE
k-NN	0,697	0,334	0,697	0,697	0.363	0.303	0.550
NB	0,767	0,257	0,767	0,767	0.510	0.276	0.406
DT-C4.5	0,757	0,276	0,756	0,756	0.486	0.281	0.452
DT-RepTree	0,772	0,268	0,770	0,770	0.513	0.296	0.407
DT-RF	0,792	0,256	0,790	0,788	0.550	0.297	0.382
ANN	0,772	0,255	0,771	0,771	0.519	0.277	0.400
SVM-POLY	0,795	0,255	0,795	0,791	0.556	**0.204**	0.452
SVM-RBF	0,781	0,299	0,789	0,770	0.511	0.219	0.468
SVM-PUK	0,783	0,271	0,781	0,778	0.528	0.217	0.466
LR	**0,798**	**0,245**	**0,797**	**0,795**	**0.566**	0.286	**0.378**

Table 5. Results of phase 4 with 9 features

Classifier	TP Rate	FP Rate	Precision	F-Measure	κ statistic	MAE	RMSE
k-NN	0,702	0,334	0,700	0,701	0.370	0.298	0.546
NB	0,766	0,285	0,763	0,762	0.494	0.293	0.410
DT-C4.5	0,777	0,270	0,775	0,774	0.520	0.291	0.419
DT-RepTree	0,771	0,274	0,769	0,768	0.508	0.299	0.409
DT-RF	0,784	0,264	0,782	0,780	0.534	0.296	0.390
ANN	0,784	0,256	0,782	0,782	0.538	0.291	0.390
SVM-POLY	0,791	0,268	**0,792**	0,785	0.543	**0.209**	0.457
SVM-RBF	0,759	0,342	0,777	0,741	0.452	0.240	0.490
SVM-PUK	**0,792**	0,265	**0,792**	**0,786**	**0.546**	**0.208**	0.456
LR	0,790	**0,259**	0,788	**0,786**	**0.546**	0.298	**0.385**

In Phase 5, the sensitivity analysis is applied to select effective features. As a result of sensitivity analysis, it is seen that BSBG03, BSBGHER, BSBGSCM and BSDMLOWP features are the most effective. The performance results calculated by using these 4 features are given in Table 6. LR is found the superior algorithm based on all selection criteria except FP rate and MAE. Also, ANN and SVM-PUK have similar results with LR.

Table 6. Results of phase 5 with 4 features

Classifier	TP Rate	FP Rate	Precision	F-Measure	κ statistic	MAE	RMSE
k-NN	0,760	0,292	0,758	0,756	0.482	0.295	0.425
NB	0,783	0,254	0,781	0,781	0.538	0.301	0.394
DT-C4.5	0,789	0,251	0,787	0,786	0.548	0.302	0.393
DT-RepTree	0,780	0,263	0,778	0,777	0.528	0.303	0.394
DT-RF	0,764	0,277	0,762	0,762	0.496	0.295	0.410
ANN	0,790	**0,250**	0,788	0,787	0.550	0.297	**0.387**
SVM-POLY	0,788	0,267	0,788	0,783	0.539	0.211	0.460
SVM-RBF	0,754	0,359	0,781	0,730	0.432	0.246	0.496
SVM-PUK	0,789	0,267	0,789	0,784	0.541	**0.210**	0.458
LR	**0,793**	0,253	**0,792**	**0,790**	**0.554**	0.302	**0.387**

The results obtained from Phase 1 to Phase 5 are summarized in Table 7. As it is seen, in all phases, LR gives the most successful classification results according to selection criteria. Also, the classification results of SVM-POLY and SVM-PUK are very close to LR, respectively. In addition to this, Table 7 shows the effective features that draw attention in the feature selection methods implemented in Phase 2 to Phase 5. BSBGHER, BSBGSCM and BSDMLOWP features are common in all Phases. Therefore, it can be said that these features are the most important features in classifying the success of the students.

Table 7. Summarized results of phase 1 to phase 5

	Algorithms	**Features**
Phase 1	LR, DT-RF, SVM-POLY	All features
Phase 2	DT-C4.5, SVM-POLY, LR	BSBG08, BSBGHER, BSBGSCM, BSDMLOWP
Phase 3	LR, SVM-POLY	BSBG06F, BSBG06K, BSBG08, BSBG13A, BSBG14B, BSBG14C, BSBGHER, BSBGSSB, BSBGSB, BSBGSLM, BSBGEML, BSBGSCM, BSBGSVM, BSDMLOWP, BSDMWKHW
Phase 4	LR, SVM-POLY,SVM-PUK	BSBGHER, BSBGSSB, BSBGSB, BSBGSLM, BSBGEML, BSBGSCM, BSBGSVM, BSDMLOWP, BSDMWKHW
Phase 5	LR, ANN, SVM-PUK	BSBG03, BSBGHER, BSBGSCM, BSDMLOWP

To investigate the performance changes of classification algorithms, one of the classification criteria can be used. In this study, TP Rate is chosen to show changes. As it is mentioned before, Table 2 shows the performances of classification algorithm with using all 35 features (Phase 1). In this Phase, the highest TP Rates are belong to LR (80,3%), DT-RF (80,5%) and SVM-POLY (80,4%) and these classification accuracies are the highest values among Phase 1 to Phase 5. When the classification accuracies (according to TP rate) of Phase 1 and other Phases are compared, the following results are obtained. In Phase 2, with the 4 most effective features, TP rates are found as DT-C4.5 (78,9%), LR (79%) and SVM-POLY (79,1%). In Phase 3, there are 15 important features and with these features, the TP rates are LR (79,8%) and SVM-POLY (79,5%). In Phase 4, with 4 extracted features, the TP rates are found as SVM-PUK (79,2%), SVM-POLY (79,1%) and LR (79%). Lastly, in Phase 5, with 4 most effective features, the TP rates are found as LR (79,3%), ANN (79%) and SVM-PUK (78,9%). According to TP rates of Phase 2 to Phase 5, it is seen that the values of classification accuracies of algorithms that are mentioned above are nearly the same when compared with Phase 1. This shows the importance of reducing features, in other words, the importance of parsimony principle. There is no significant difference between the result with using all features and the result with using most important features.

5. DISCUSSION AND CONCLUSION

It is very important to determine the determinants/features of students' success in educational decision-making process. As the features affecting the success of the students are known, it can be assumed that the success of the learner will be increased by taking precautions for these features or changing the current conditions. Based on this idea, the current study focuses on two research questions: first, which algorithms are appropriate to classify students' success and, second, what are the most important features in classification with using appropriate algorithms. In order to address these research questions, TIMSS 2015 8th grade Turkish students' mathematics data is handled.

For the first research question, most of algorithms that have been performed in many studies within the concept of EDM are used. The findings show that LR is the most appropriate

algorithms according to many selection criteria. Besides this, the performance of SVM-POLY and SVM-PUK are nearly similar with LR. When the literature is examined, it is seen that many important studies claim that LR is a useful tool in modeling educational data. Schreiber [59] found the LR as the best modeling method in order to examine the important features that affect the academic performance. Besides, in the study of Yoo [60], TIMSS 2011 Korean 4th grade students' mathematics scores are taken as the target population. LR was employed to data as prediction method and results were reported by performing this most popular machine learning technique. Another study which found the LR as the best performing algorithm among different data mining methods was done by Kılıç-Depren, Askin and Öz [22]. Delen [61] used different data mining methods such as C5 decision tree algorithm, ANN, SVM and LR to construct analytical models for predicting freshmen students' attrition. The obtained results showed that SVM gives the better performance. In the study of Gorostiaga and Rojo-Álvarez [62], LR, Fisher discriminant analysis and SVM (both linear and non-linear) are used to classify PISA2009 mathematics result of Spanish students' academic scores. It was seen that SVM outperformed than other algorithms.

The second research question is more likely to make correct identification in students' mathematics success. For this purpose, reducing procedure is implemented to all features and the effect of reducing the number of features to the change the classification performance is investigated with the help of the different feature selection methods. Accordingly, it is determined that different feature selection methods give different classification results. Having the loss as result of the reduction in the number of features less than the information gain shows that the study gives significant results. As a result of this procedure, "Home Educational Resources", "Student Confident in Mathematics" and "Mathematics Achievement Too Low for Estimation" are common in results of the all feature selection algorithms. When the studies in the literature are examined, it is seen that the results obtained in the context of using different analyzes/methods in each study support the results of this study. Liu and Meng [19], Hammouri [18], and Askin and Gokalp [20], Kılıç-Depren, Askin and Öz [22] found in their studies that "Student Confident" is an important factor on academic achievement. In addition, "Home educational resources" is found as another important feature in success and in the study of Topçu, Erbilgin & Arıkan [21], it is pointed out that students who can easily reach the educational resources often become successful.

This study contributes the existing literature by different perspectives. First of all, TIMSS 2015 is the last released data for mathematics and science achievement and there are few studies deal with this latest released data. Secondly, when the entire TIMSS literature has been investigated in terms of methodology, it can be seen that many of them preferred standard statistical methods in order to perform clustering, prediction and regression. However, because of the strict assumptions of standard statistical methods, data mining algorithms has been becoming very popular. Data mining algorithms performed in this study are the most selected and used algorithms when the EDM literature is reviewed. Thirdly, the current study underlines the importance of data reduction in the process of classification of students' achievement. There is no need to include all features into the model, in other words, only most important features can be used in order to make successful classification. In addition to these contributions, educators and education policy makers can be used the most important features extracted in this study. To improve the students' academic achievement, knowing important features plays a vital role. For future studies, further investigations can be done for science results for TIMSS 2015. These EDM algorithms can be applied to different national/international studies such as PISA to make classification of academic achievement.

TIMSS 2015 dataset used in this study is based on the assessment which shows the results of self-reported questionnaires prepared by IEA, and this is one of the limitations of current study. Besides, the other limitation is that findings of this study are only for TIMSS 2015 8th grade students' data and only for their mathematics successes.

Acknowledgments

This paper was derived from the doctoral dissertation by Enes Filiz conducted under the supervision of Assoc. Prof. Dr. Ersoy Öz.

REFERENCES

[1] Mullis, I.V., Martin, M.O., Foy, P., Arora, A., (2012) *TIMSS 2011 international results in mathematics*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

[2] Han, J., Kamber, M., Pei, J., (2012) *Data mining: Concept and techniques*, (3rd ed.). MA: Morgan Kaufmann Publishers, Burlington.

[3] Sinharay, S., (2016) An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice* 35, 38-54.

[4] Ramaswami, M., Bhaskaran, R., (2012) A CHAID based performance prediction model in educational data mining. *arXiv preprint arXiv:1002.1144*.

[5] Romero, C., Ventura, S., (2010) Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40, 601-618.

[6] Romero, C., Ventura, S., (2007) Educational data mining: A survey from 1995 to 2005. *Expert systems with applications* 33, 135-146.

[7] Baker, R.S., Yacef, K., (2009) The state of educational data mining in 2009: A review and future visions. *JEDM Journal of Educational Data Mining* 1, 3-17.

[8] Siemens, G., Baker, R.S., (2012) Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge, 2012, April, ACM.

[9] Mohamad, S.K., Tasir, Z., (2013) Educational data mining: A review. *Procedia-Social and Behavioral Sciences* 97, 320-324.

[10] Peña-Ayala, A., (2014) Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications* 41, 1432-1462.

[11] Shahiri, A.M., Husain, W., (2015) A review on predicting student's performance using data mining techniques. *Procedia Computer Science* 72, 414-422.

[12] Kotsiantis, S., Pierrakeas, C., Pintelas, P., (2004) Predicting Students' performance in Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence* 18, 411-426.

[13] Baradwaj, B.K., Pal, S., (2011) Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications* 2, 63-69.

[14] Rajni, J., Malaya, D.B., (2015) Predictive analytics in a higher education context. *IT Professional* 17, 24-33.

[15] Martínez Abad, F., Chaparro Caso López, A.A., (2017) Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement* 28, 39-55.

[16] Cortez, P., Silva, A.M.G., (2008) Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology. Conference. Porto, Portugal: EUROSIS, 5-12.

[17] Osmanbegović, E., Suljić, M., (2012) Data mining approach for predicting student performance. *Economic Review* 10, 3-12.

[18] Hammouri, H., (2010) Attitudinal and motivational variables related to mathematics achievement in Jordan: Findings from the Third International Mathematics and Science Study (TIMSS). *Educational Research* 46, 241-257.

[19] Liu, S., Meng, L., (2010) Re-examining factor structure of the attitudinal items from TIMSS 2003 in cross-cultural study of mathematics self-concept. *Educational Psychology* 30, 699-712.
[20] Askin, O.E., Gokalp, F., (2013) Comparing the predictive and classification performances of logistic regression and neural networks: a case study on timss 2011. *Procedia-Social and Behavioral Sciences* 106, 667-676.
[21] Topçu, M.S., Erbilgin, E., Arıkan, S., (2016) Factors Predicting Turkish and Korean Students' Science and Mathematics Achievement in TIMSS 2011. *Eurasia Journal of Mathematics, Science & Technology Education* 12, 1711-1737.
[22] Kılıç-Depren, S., Askin, Ö.E., Öz, E., (2017) Identifying the Classification Performances of Educational Data Mining Methods: A Case Study for TIMSS. *Educational Sciences: Theory & Practice* 17, 1605-1623.
[23] Filiz, E., Oz, E., (2019) Finding The Best Algorithms and Effective Factors in Classification of Turkish Science Student Success. *Journal of Baltic Science Education* 18, 239-253.
[24] LaRoche, S., Joncas, M., Foy, P., (2016) Sample Design in TIMSS 2015. Martin, M. O., Mullis, I.V.S., and Hooper, M. (Eds.), Methods and Procedures in TIMSS 2015. Retrieved from Boston College, TIMSS & PIRLS International Study Center.
[25] Masters, G.N., Wright, B.D., (1997) The partial credit model. In M.J. van de Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. Berlin: Springer.
[26] Filiz, E., Öz, E., (2017) Classification of BIST-100 Index' Changes via Machine Learning Methods. *Marmara University Journal of Economic & Administrative Sciences* 39, 117-129.
[27] Jiang, S., Pang, G., Wu, M., Kuang, L., (2012) An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications* 39, 1503-1509.
[28] Li, B., Yu, S., Lu, Q., (2003) An improved k-nearest neighbor algorithm for text categorization, *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages*, 3-6 August 2003, Shenyang, China.
[29] John, G.H., Langley, P., (1995) Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, August, Morgan Kaufmann Publishers Inc.
[30] Zhang, H., (2004) The optimality of naive Bayes. *AA* 1, 3.
[31] Haykin, S., (1999) Neural Networks: A comprehensive Foundation, *Prentice Hall International*. Inc., Englewood Cliffs.
[32] Alpaydın, E., (2004) *Introduction to machine learning*. MIT press, Cambridge.
[33] Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M., (1998) Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory* 44, 1926-1940.
[34] Varshney, P.K., Arora, M.K., (2004) *Advanced image processing techniques for remotely sensed hyperspectral data*. Springer Science & Business Media.
[35] Rokach, L., Maimon, O., (2005) Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165-192) Springer, Boston, MA.
[36] Quinlan, J.R., (2014) *C4.5: programs for machine learning*. Elsevier.
[37] Kalmegh, S., (2015) Analysis of WEKA data mining algorithm REPTree, Simple CART and RandomTree for classification of Indian news. *Int. J. Innov. Sci. Eng. Technol.* 2, 438-446.
[38] Srinivasan, D.B., Mekala, P., (2014) Mining Social Networking Data for Classification Using REPTree. *International Journal of Advance Research in Computer Science and Management Studies* 2, 155-160.
[39] Chen, X.W., Liu, M., (2005) Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21, 4394-4400.

[40] Breiman, L., (2001) Random forests. *Machine learning* 45, 5-32.
[41] Hosmer, D.W., Lemeshow, S., (2000) *Applied Logistic Regression*, 2nd ed.; Hoboken, NJ: John Wiley & Sons, Inc.
[42] Balaban, M.E., Kartal, E., (2015) *Basic Algorithms of Data Mining and Machine Learning and Applications with R Language*. Çağlayan Kitabevi, İstanbul.
[43] Donner, A., Klar, N., (1996) The statistical analysis of kappa statistics in multiple samples. *Journal of clinical epidemiology* 49, 1053-1058.
[44] Turanoğlu-Bekar, E., Ulutagay, G., Kantarcı-Savas, S., (2016) Classification of thyroid disease by using data mining models: A comparison of decision tree algorithms. *Oxford Journal of Intelligent Decision and Data Sciences* 2, 13-28.
[45] Willmott, C.J., Matsuura, K., (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 79-82.
[46] Gennari, J.H., Langley, P., Fisher, D., (1989) Models of incremental concept formation. *Artificial intelligence* 40, 11-61.
[47] Gümüşçü, A., Aydilek, İ.B., Taşaltın, R., (2016) Comparison of Feature Selection Algorithms on Microarray Data Classification. *Harran University Journal of Engineering* 1, 1-7.
[48] Kononenko, I., (1994) Estimating attributes: analysis and extensions of RELIEF, In *European conference on machine learning*, Springer: Berlin, Heidelberg.
[49] Cover, T.M., Thomas, J.A., (2012) Elements of information theory. John Wiley & Sons.
[50] Aktas, M.S., Kalıpsız, O., (2015) Research and Comparative Practice on the Application of Feature Selection Techniques to Banking Data in Data Mining. In Proceedings of the 9th Turkish National Software Engineering Symposium, 2015, September; Yasar University: Izmir, Turkey.
[51] Muda, Z., Yassin, W., Sulaiman, M.N., Udzir, N.I., (2011) Intrusion detection based on k-means clustering and OneR classification, *7th International Conference on Information Assurance and Security IAS2011*, 5-8 December 2011, Malacca, Malaysia.
[52] Kabakchieva, D., (2013) Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies* 13, 61-72.
[53] Saltelli, A., Chan, K., Scott, E.M., (2000) *Sensitivity analysis* (Vol. 1). New York: Wiley.
[54] Gondra, I., (2008) Applying machine learning to software fault-proneness prediction. *Journal of Systems and Software* 81, 186-195.
[55] Yao, J.T., (2003) Sensitivity analysis for data mining, *NAFIPS 2003 22nd International Conference of the North American Fuzzy Information Processing Society*, 24-26 July 2003, Chicago, IL, USA.
[56] Gazi, V.E., (2007) Data Mining Sensitivity. Master's thesis, Institute of Science and Technology Istanbul Technical University, Turkey.
[57] Hall, M.A., (2000) Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, The University of Waikato, Working Paper 00/8. Hamilton, New Zealand.
[58] Frank, E., Mark, A.H., Ian, H.W., (2016) *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Fourth Edition; Morgan Kaufmann.
[59] Schreiber, J.B., (2002) Scoring Above the International Average: A Logistic Regression Model of the TIMSS Advanced Mathematics Exam. *Multiple Linear Regression Viewpoints* 28, 22-30.
[60] Yoo J.E., (2018) TIMSS 2011 Student and Teacher Predictors for Mathematics Achievement Explored and Identified via Elastic Net. *Frontiers in psychology* 9, 317.
[61] Delen, D. A., (2010) comparative analysis of machine learning techniques for student retention management. *Decision Support Systems* 49, 498-506.