KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: December 13, 2016
Revision received: May 22, 2017
Accepted: June 7, 2017
OnlineFirst: August 10, 2017

www.estp.com.tr
DOI 10.12738/estp.2017.5.0634 ♦ October 2017 ♦ 17(5) ♦ 1605–1623

Research Article

Identifying the Classification Performances of Educational Data Mining Methods: A Case Study for TIMSS

Serpil Kılıç Depren[1]
Yildiz Technical University

Öyküm Esra Aşkın[2]
Yildiz Technical University

Ersoy Öz[3]
Yildiz Technical University

Abstract

Educational data mining (EDM) is a rapidly growing research area, and the outputs obtained from EDM shed light on educators' and education planners' efforts to make efficient decisions concerning educational strategies. However, a lack of work still exists on using EDM methods for international assessment studies such as the International Association for the Evaluation of Educational Achievement's Trends in International Mathematics and Science Study (IEA's TIMSS). This study aims to fill the gap in the current literature on the latest-released TIMSS 2011 data by applying a decision tree, a Bayesian network, a logistic regression, and neural networks. The best performing algorithm in classification based on several performance measures has been found for eighth-grade Turkish students' mathematics data. During the construction of models, 11 student-based factors have been taken into account. The results show that logistic regression outperforms other algorithms in terms of measuring classification performance. The factor of student confidence has also been found as the most effective factor on eighth-grade students' mathematics achievement.

Keywords

Classification algorithms • Classification performance measures • Educational data mining • Mathematics achievement • TIMSS

[1] **Correspondence to**: Serpil Kılıç Depren (PhD), Department of Statistics, Faculty of Arts and Science, Yildiz Technical University, Davutpaşa Campus, Esenler, Istanbul 34220 Turkey. Email: serkilic@yildiz.edu.tr

[2] Department of Statistics, Faculty of Arts and Science, Yildiz Technical University, Esenler, Istanbul 34220 Turkey. Email: oeyigit@yildiz.edu.tr

[3] Department of Statistics, Faculty of Arts and Science, Yildiz Technical University, Esenler, Istanbul 34220 Turkey. Email: ersoyoz@yildiz.edu.tr

Citation: Kılıç Depren, S., Aşkın, 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. http://dx.doi.org/10.12738/estp.2017.5.0634

Data mining is a multidisciplinary study area that includes many different statistical procedures. The main goal of data mining is to explore useful hidden patterns among huge data sets. Unlike traditional statistical methods such as linear regression, data mining methods do not require the assumptions of linearity, variance, homogeneity, or normality (Sinharay, 2016).

With the rapid development of information technologies, a great number of techniques within data mining have been applied over many disciplines, including social sciences, physics, engineering, and medicine. Studies that use data mining methods in the field of education are generally known as educational data mining (EDM) studies, of which an extensive literature review was performed by Romero and Ventura (2007) for the period between 1995 and 2005. In their study, EDM is stated as an iterative cycle that includes hypothesis formation, testing, and refinement. They pointed out that the outputs obtained from EDM methods guide educators in their discovery of useful information on formative evaluation. Through the usefulness of this newly discovered information, educators have established a pedagogical basis for decisions when designing or modifying an environment or teaching approach (Romero & Ventura, 2007). Another important study by Baker and Yacef (2009) dealt with EDM and its major trends. Their study discussed four different domains within the field of EDM, addressing studies with regard to each of these domains. In addition to these substantial works, Pena-Ayala's (2014) study provided an EDM survey from the beginning of 2010 until the first quarter of 2013 that included 240 published papers. Despite the important studies that exist in the literature on EDM mining, it still lacks research, particularly on supervised learning methods for huge data sets that require computationally difficult algorithms (Sinharay, 2016). One of the biggest known challenges facing education planners is how to analyze huge data sets in terms of student's characteristics such as knowledge, motivation, and attitudes (Baker, 2010). In the process of improving the quality of managerial decisions for future education strategies, understanding and exploring the hidden patterns from observable data is generally difficult and time consuming when done manually (Mohamad & Tasir, 2013). Therefore, much attention should be given to the study of EMD in order to enlighten educators and education planners. Output obtained from EDM can offer need-oriented solutions through different perspectives in the process of determining useful education strategies (Bilen, Hotaman, Aşkın, & Büyüklü, 2014).

Several studies that fall within the concept of EDM have been carried out in the recent literature in order to provide reliable solutions regarding educational phenomena. He's (2013) study pointed out that assessing learning performance, providing feedback, and adapting learning materials based on students' learning behaviors are some of the reasons to use EDM. Cortez and Silva's (2008) study applied four supervised learning methods (i.e., decision tree, random forest, neural networks, support vector machines) for building a model based on the student performance of secondary schools in Portugal. Their findings showed not only the best prediction model but also supported the idea that academic achievement and past performance highly correlate with each other. Ramaswami and Bhaskaran (2010) collected data from higher education students and constructed a prediction model based on CHAID, which is the most commonly used decision tree algorithm. Furthermore, their results identified statistically significant factors influencing academic performance. The studies of Alivernini (2013), Abad and Lopez (2017), and İdil, Narlı, and Aksoy (2016) also dealt with decision trees for constructing student-based models that would ensure accurate classifications and predictions, but respectively used the different algorithms of CART, J48 and C5.0. Kotsiantis, Pierrakeas, and Pintelas (2010) investigated the prediction performance of six different machine-learning methods/algorithms (decision tree/C4.5, neural networks/back propagation, Bayesian network/naive Bayes, instance-based learning/k-nearest neighbor, logistic regression/maximum likelihood, and support vector machines/sequential minimal optimization). Naive Bayes was found to be the most appropriate algorithm for predicting the performance of students registered in a distance-learning program according to the different criteria used in the study. The purpose of Vialardi et al.'s (2011) study was to present a recommendation system based on the records of university students' academic performance. This system helps students make proper decisions during the enrollment process. Ensemble-learning approaches such as bagging and boosting, tree based algorithms such as C4.5 and k-nearest neighbor, and the naive Bayes algorithm were performed, showing that bagging provides the best prediction accuracy among these at 85.36%. In the studies of Ramesh, Parkavi, and Ramar (2013) and Shahiri, Husain, and Rashid (2015), neural networks were found to be the best performing algorithm for predicting students' academic performance when compared to decision trees, naive Bayes, and support-vector machines.

TIMSS Literature

Despite the several studies done nationally in Turkey regarding EDM, a lack of work still exists on using supervised learning methods for international assessment studies. IES' Trends in International Mathematics and Science Study (TIMSS), conducted every four years, deals with mathematics and science students in the fourth and eighth grades. TIMSS not only provides information about the effects of policies and practices in each participating country's education system (Mullis, Martin, Foy, & Arora, 2012), it also enables researchers to make comparisons among the results in terms of student achievements. Standard statistical procedures such as regression analysis, multilevel modeling, and factor analysis have been widely applied using students' characteristics as influencing factors when investigating science and mathematics achievement. (Kilic & Askin, 2013; Neuschmidt, Barth, &

Hastedt, 2008; Schreiber, 2002; Topçu, Erbilgin, & Arıkan, 2016; Wößmann, 2005). As mentioned previously, however, these methods have some drawbacks, especially with abnormally distributed data. Furthermore, when the problem is relatively complex, difficulties arise in obtaining accurate predictions (Razi & Athappilly, 2005). Supervised learning algorithms are robust at identifying outliers (Dejaeger, Goethals, Giangreco, Mola, & Baesens, 2012) and can be used to overcome certain limitations in these standard procedures.

The Present Study

This study aims to fill the gap on TIMSS 2011 (the latest-released data) in the current literature by performing a decision tree, Bayesian network, logistic regression, and neural networks. Different algorithms within the context of these methods are compared in terms of their classification accuracy. In this process, widely used performance measures have been taken into account. While selecting the best-performing classification algorithm, Turkish eighth-grade students' characteristics (i.e., age, gender, family background) have been included in models as potential influencing factors on mathematics achievement. Furthermore, Friedman's two-way analysis of variance is performed in order to show whether the differences among algorithms are statistically significant. In this context, the study's research questions have been designed as follows:

1. Which method has the best classification performance based on model-performance measures?
2. Which factors are found significant on mathematics achievement; what is their order of importance?
3. What measures should be taken to improve Turkish eighth-graders' mathematics achievement?

Data Set and Factors

Data source. TIMSS started its first assessment in 1995, and has been conducted regularly every four years since by the International Association for the Evaluation of Educational Achievement. TIMSS 2011 contains the mathematics achievement results of 42 countries and 14 benchmarking participants from the 8th grade. The data analyzed in this study comes from TIMSS 2011's eighth-grade Turkish students. A total of 6,928 students (3,414 females and 3,514 males) have been sampled. Due to some missing and inaccurate values from the original data, 678 students' data sets have been excluded, leaving a total of 6,250 data sets.

Factors. Mathematics assessment is designed along two dimensions of skills: content (numbers, algebra, geometry, data, and probability) and cognitive (knowing, applying, and reasoning; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). TIMSS scores are scaled to have an international average value of 500 and a standard deviation of 100 points across all countries (Mullis et al., 2012). The average scale score in mathematics for Turkey (452) is below the TIMSS scale's central score (500). Note here that the first plausible value is chosen as the response variable *MATHACH*, and if the score is greater than the TIMSS scale's central score, then *MATHACH* = 1; if not, *MATHACH* = 0. The description of factors that potentially influence the *MATHACH* score is given in Table 1.

Table 1
Student Related Factors

Factor Name	Description	Domain
AGE	Student age	Scale measurement (age 13 through 16)
SEX	Student gender	1 = Female, 2 = Male
HER	Home educational resources	1 = Many resources, 2 = Some resources, 3 = Few resources
SB	Student bullied at school	1 = Almost never, 2 = About monthly, 3 = About weekly
SLL	Students like learning math	1 = Like learning mathematics, 2 = Somewhat like learning mathematics, 3 = Don't like learning mathematics
SVL	Students value math	1 = Value, 2 = Somewhat value, 3 = Don't value
SC	Students confident in math	1 = Confident, 2 = Somewhat confident, 3 = Not confident
SE	Students engaged in math	1 = Engaged, 2 = Somewhat engaged, 3 = Not engaged
HSS	Number of home study supports	0 = Neither own room nor internet connection, 1 = Either own room or internet connection, 2 = Both own room and internet connection
PHEL	Parents' highest education level	1 = University or higher, 2 = Post-secondary but not university, 3 = Upper secondary, 4 = Lower secondary, 5 = Some primary, lower secondary or no school
WTSMH	Weekly time spent on math homework	1 = 3 Hours or more, 2 = More than 45 minutes but less than 3 hours, 3 = 45 minutes or less
MATHACH (Response Variable)	Whether successful or not	0: Not successful, 1: Successful

Method

Classification Algorithms

Two decision tree algorithms (random forest and J48), a Bayesian network algorithm (naive Bayes), an artificial neural-networks algorithm (multilayer perceptron), and the logistic regression algorithm will be introduced in this section. These algorithms are often preferred by researchers due to their classification successes.

Random Forest (RF). RF has been widely used in classification problems and can be simply described as a combination of tree-structured classifiers (Breiman, 2001). For the k^{th} (k = 1,...) tree, an independent identically distributed random vector Θ_k.is sampled from the past vectors which are denoted by Θ_1 ,..., Θ_{k-1}. Here, the past random vectors have the same distribution with Θ_k. If the input vector is shown with , a tree is constructed using the training set and the random vector Θ_k. When large numbers of tree are generated, each tree votes for a unit (in other words produce a classification). The result can be defined as $h(x, \Theta_k)$, which has the most votes collected from the trees (Breiman, 2001; Turanoğlu-Bekar, Ulutagay, & Kantarcı-Savaş, 2016).

J48. J48 is a decision-tree algorithm that uses an open-source Java implantation with a revised version of the C4.5 algorithm (Quinlan, 1993). This algorithm was developed by Quinlan (1993) to overcome the deficiencies and inadequacies of the ID3 algorithm. It has the ability of handling missing values in the training data set by predicting them from the observable attributes. It can also be implemented in datasets that include both discrete and continuous attributes. While building a tree, a pruning method is used in order to reduce the tree size by removing over-fitting data; the data is classified recursively until the best categorization has been achieved (Dangare & Apte, 2012). The algorithm generates trees using information gain and entropy. Here, information gain is a quantity that describes how well the given attribute is distinguished from the training set. At each decision node, the most useful attribute, which means the attribute with the highest information gain, is selected. Let a sample space (training data set) be shown by S, and the relevant information gain be denoted by $Gain_{(S,\ A)}$. Then, the information gain can be calculated as follows (Sugumaran, Muralidharan, & Ramachandran, 2007):

$$Gain_{(S,A)} = Entropy_{(S)} - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy_{(S_v)} \tag{1}$$

where all possible values of the attribute A are denoted by Value(A), and S_v is a subset of sample space S $(S_v = \{s \in S \,|\, A(s) = v\})$. It should be noted that homogeneity is measured by entropy, given in the equation:

$$Entropy_{(S)} = -\sum_{i=1}^{c} P_i \log_2 P_i \tag{2}$$

where P_{i} is the probability that S belongs to the i^{th} class and c is the number of classes.

Naive Bayes (NB). The NB classifier can be described as a special form of Bayesian networks. The algorithm got its name "naive" because two assumptions need to be met. Firstly, factors should be conditionally independent with respect to class. Secondly, factors affecting the interested outcome are assumed to not be hidden; in other words, no latent factors influence the prediction (John & Langley, 1995). When the assumption of independency is satisfied, the learning process of Bayesian classifiers becomes simpler, and optimal assignment is achieved using the vector of observable factors (Öz, Kurt, Asyali, Kaya, & Yucel, 2016). Let $X = (X1,...,X_n)$ show the vector of observable factors and the random variable C denote a class. Then the NB classifier can be written as:

$$P(X|C) = \prod_{i=1}^{n} Prob(X_i|C) \tag{3}$$

Multilayer perceptron. The main idea of artificial neural networks (ANNs) is for them to mimic the processes of the human brain; these models can learn and generalize from past experience by training them like a human brain. The main advantage of using ANNs is that no priori assumptions need to be met as in standard statistical methods, even when dealing with complex nonlinear relations. The multilayer perceptron (MLP) is an ANN model that uses a back-propagation algorithm in the training process. MLP has three components: an input layer, hidden layers, and an output layer. Information is carried from one neuron to another by the weight value. The first step of an MLP algorithm is to randomly assign weights. In the second step, the inputs (independent variables) propagate forward using the sigmoid or logistic-activation function, thus producing output values (dependent variables) for each hidden layer. After that, the error is propagated backward by updating the weights and biases from Step 3. Errors are computed for each output and hidden layer. In the final step, weights and biases are updated and returned to Step 2. The steps are repeated until the overall error is minimized (Han, Kamber, & Pei, 2006).

Logistic regression. As in standard regression models, the relationship between dependent and independent variables is investigated using logistic regression. However, a main assumption of standard regression is that the dependent variable (generally shown as Y) needs to be continuous. When Y takes a value of 0 or 1, binary logistic regression is performed in order to predict Y from the observable independent variable X. The simplest binary logistic regression model is given as:

$$\pi(x) = \frac{e^{x'ß}}{1-e^{x'ß}} = \frac{1}{e^{x'ß}} \qquad (4)$$

where $\pi(x) = E[Y/X = x]$ and β is a vector of the k regression parameters. A logit transformation is applied because the model shown by Equation 4 is nonlinear.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = x'ß \qquad (5)$$

As seen in Equation 5, logit $\pi(x)$ is the natural logarithm of odds and is linear. The unknown regression coefficients $\beta^t = (\beta_0, \beta_1,..., \beta_k)$are estimated using the maximum likelihood estimation method, which maximizes the logarithmic-likelihood function (Hosmer & Lemeshow, 2000).

Classification Performance Measures

Different performance measures are implied to assess the classifiers, and these measures are evaluated together to produce more accurate results. Commonly used measures have been chosen and provided in this section.

Kappa statistics. Kappa statistics (κ), a goodness-of-fit statistic for measuring inter-rater agreement among categorical variables, evaluates the prediction performance of a classification model and is based on the chi-square table (Donner & Klar, 1996). The closer κ is to 1, the higher the agreement between raters. Let P_0 and P_e denote the observed agreement between two categorical variables and change-expected agreement, respectively (Turanoğlu-Bekar et al., 2016). Then, κ can be obtained using:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{6}$$

***MAE* and *RMSE* statistics**. Mean absolute error (*MAE*) and root-mean-square error (*RMSE*) statistics describe the differences between a model's predicted and observed values (Willmott & Matsuura, 2005). *MAE* measures the average of the absolute differences between predicted and observed values. These differences have equal weight. *RMSE* is the square root of the average of the squared differences between predicted and observed values. The relation between *MAE* and *RMSE* can be written as $MAE \leq RMSE$ and these statistics are calculated as follows:

$$MAE = n^{-1} \sum_{i=1}^{n} |P_i - O_i| \tag{7}$$

$$RMSE = \sqrt{n^{-1} \sum_{i=1}^{n} |P_i - O_i|^2} \tag{8}$$

where P_i and O_i ($i = 1,...,n$) are the predicted and observed values, respectively. Here, the values of $P_i - O_i$ represents the model prediction errors.

TP* rate, *FP* rate, precision, and *MCC. The true positive (*TP*) rate, sometimes called *sensitivity*, is the proportion of positives that are classified correctly. The true negative (*TN*) rate, also called *specificity*, denotes the number of correctly classified negative samples. A false positive (*FP*, or Type-I error) is when the null hypothesis is actually true; the hypothesis test is declared significant in this condition. False negative (*FN*, or Type -II error) is the error of not rejecting the null hypothesis when it is actually false. Precision is the proportion of positively classified samples that are indeed positive. The Matthews correlation coefficient (*MCC*) takes values between -1 and 1 and is obtained by using elements from the confusion matrix. *MCC*s with a positive value can conclude that correct predictions have been derived (*MCC* = 1 means perfect predictions). Formulas for these measures are given as:

$$\text{TP rate} = \text{TP} / (\text{TP} + \text{FN}) \tag{9}$$

$$\text{TN rate} = \text{TN} / (\text{TN} + \text{FP}) \tag{10}$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{11}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{12}$$

***F*-measure**. The *f*-measure is calculated by taking a weighted average of the precision and *TP* rate, and can be obtained by:

$$f\text{-}measure = \frac{(1+ß^2)\times precision\times TP}{(ß^2\times precision+TP)} \tag{13}$$

where *ß* is the relative-importance value of the precision and TP rate. In general, this value is taken as 1.

***ROC* Area and *PRC* Area**. The area under the Receiver Operating Characteristic (*ROC*) curve is an illustration of how a classifier performs and has been widely used in evaluating the performance of classification algorithms (Bradley, 1997). The *ROC* area-curve is drawn with the *TP* value on the *Y*-axis and (1 - *TN*) value on the *X*-axis at every possible threshold. A higher *ROC*-area value suggests better classification to have been achieved by the related algorithm. The precision-recall curve (*PRC*) is a two dimensional graph like the *ROC* but however uses the rate of precision (Eq. 11) on the *Y*-axis and the rate of recall (Eq. 9) on the *X*-axis. As in the *ROC* curve, the area under the *PRC* curve gives an idea about the classification performances of the classifiers being compared. The main difference between the two curves in the application is the structural behavior of the dataset. Davis and Goadrich (2006) indicated that *PRC* is preferable with highly skewed data.

Experimental Setup

Step 1: Preparing the data and selecting factors. IES's TIMSS 2011, the latest dataset released for eighth-grade Turkish students, is obtained. The dataset contains information from 6,250 students. A total of 11 factors are chosen as the independent variables, and one binary variable (having a value of 0 or 1) is taken as a dependent variable to indicate mathematic achievement.

Step 2: Creating the training and testing data. Classification is performed based on the dependent variable. The dataset is split into two sets: training and testing. The system is trained using the training set and performance evaluation is done using the testing set. In this study, the common technique of 10-fold cross validation (Stone, 1974) has been used to assess algorithms' classification performances. This technique splits the data into 10 equal sets (folds). One set is used for training, and nine sets are used for testing. This process is repeated 10 times, and the average *k* result is recorded as the classification accuracy of the related classifier.

Step 3: Performing and evaluating classification algorithms. The classification accuracies (classifier performances) are obtained. The classifiers used in this study are: random forest, J48, naive Bayes, multilayer perceptron, and logistic regression. Analyses are performed using WEKA 3.7 (Waikato Environment for Knowledge Analysis) software. The processing and operating systems of the computer used for computations are an Intel Core i5-2400 (3.10GHz) and Windows 10 Pro 64-bit, respectively, with an additional 8GB of DDRIII-RAM. The algorithms' classification results are assessed based on the different previously mentioned performance measures. The best classification algorithm is determined in accordance with these performance measures, and factors' ranks are reported in terms of their significance to mathematics achievement.

Step 4: Testing the differences in performance measures. In step 3, the best performing algorithm is decided according to the quantities of performance measures. In order to test whether these performance measures statistically differ among the algorithms, Friedman's two-way analysis of variance is carried out.

Findings and Results

Results of Descriptive Statistics

Frequencies and percentages for student-related categorical variables are given in Table 2. According to Table 2, 50.6% of eighth-grade students' data used in this study are from females; almost half of the parents' highest education level has been flagged as "some primary, lower secondary, or none." Also, 32.2% of students have their own room and internet connection.

Table 2
Frequencies and Percentages for Student-Related Factors

Domain		1	2	3	4	5	Total
Parents' Highest Education Level	%	7.7%	4.8%	23.9%	14.5%	49.1%	100%
	n	482	297	1,494	908	3,069	6,250
Domain		1	2	3	Total		
Home educational resources	%	3.8%	41.7%	54.5%	100%		
	n	240	2,604	3,406	6,250		
Student bullied at school	%	52.5%	32.6%	14.9%	100%		
	n	3,280	2,037	933	6,250		
Students like learning math	%	31.9%	41.8%	26.3%	100%		
	n	1,993	2,614	1,643	6,250		
Students value math	%	46.1%	39.5%	14.4%	100%		
	n	2,881	2,468	901	6,250		
Students confident in math	%	14.2%	36.4%	49.5%	100%		
	n	886	2,273	3,091	6,250		
Students engaged in math	%	29.0%	59.0%	11.9%	100%		
	n	1,814	3,690	746	6,250		
Weekly time spent on math homework	%	8.2%	40.8%	51.0%	100%		
	n	510	2,550	3,190	6,250		
Domain		0	1	2	Total		
Number of home study supports	%	31.7%	36.1%	32.2%	100%		
	n	1,983	2,256	2,011	6,250		
Domain		1	2	Total			
Student's gender	%	50.6%	49.4%	100%			
	n	3,161	3,089	6,250			
Domain		0	1	Total			
MATHACH	%	67.3%	32.7%	100%			
	n	4,209	2,041	6,250			

In this study, 96.2% of eighth-grade students have some/few educational resources. Furthermore, 73.7% of them like learning math, and 46.1% think that the information they learn is valuable. Almost half do not feel confident and spend less than 45 minutes a week on their mathematics homework.

Results of Classification Algorithms

The performance of classification algorithms are compared with classification performance measures for each algorithm.

Table 3

Summary of Performance Measures for Different Classification Algorithms

	Random Forest	J48	Naive Bayes	Multilayer Perceptron	Logistic Regression
κ statistic	0.3981	0.4668	0.4669	0.4710	0.4885
MAE	0.3004	0.3072	0.2705	0.2873	0.3029
RMSE	0.4344	0.4062	0.4111	0.3981	0.3885

According to Table 3, naive Bayes, J48 and multilayer perceptron algorithms have similar kappa statistics (κ ≈ 0.47). The random forest algorithm has the lowest kappa value, while logistics regression is the best performer algorithm in terms of kappa and *RMSE* statistics, κ = 0.4885 and κ = 0.3885, respectively.

The calculated values for *TP* rate, *FP* rate, precision, recall, *f*-measure, *MCC*, *ROC*, and *PRC* for all algorithms are given in Table 4.

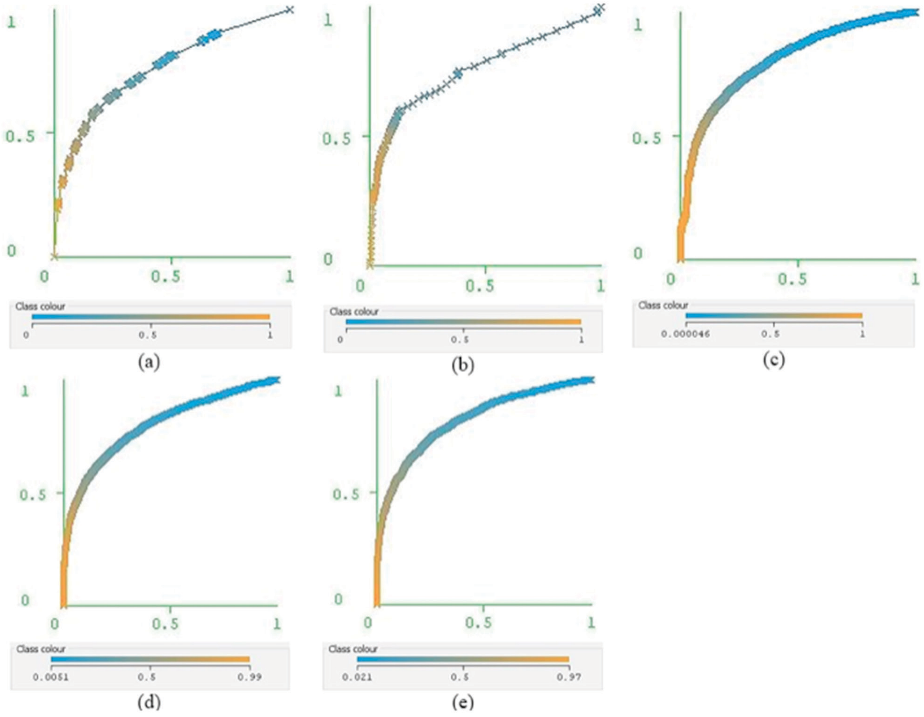
Table 4

Statistical Analysis of Algorithms

	Class	*TP* Rate	*FP* Rate	Precision	Recall	*f*-Measure	*MCC*	*ROC* Area	*PRC* Area
Random Forest	0	0.857	0.478	0.787	0.857	0.821	0.402	0.752	0.826
	1	0.522	0.143	0.640	0.522	0.575	0.402	0.752	0.636
	Weighted Avg.	0.748	0.369	0.739	0.748	0.740	0.402	0.752	0.764
J48	0	0.909	0.478	0.797	0.909	0.850	0.480	0.758	0.810
	1	0.522	0.091	0.737	0.522	0.611	0.480	0.758	0.658
	Weighted Avg	0.783	0.351	0.777	0.783	0.772	0.480	0.758	0.760
Naive Bayes	0	0.845	0.387	0.818	0.845	0.832	0.467	0.806	0.873
	1	0.613	0.155	0.658	0.613	0.635	0.467	0.806	0.722
	Weighted Avg	0.770	0.311	0.766	0.770	0.767	0.467	0.806	0.824
Multilayer Perceptron	0	0.893	0.449	0.804	0.893	0.846	0.479	0.801	0.868
	1	0.551	0.107	0.713	0.551	0.622	0.479	0.801	0.716
	Weighted Avg	0.781	0.338	0.774	0.781	0.773	0.479	0.801	0.819
Logistic Regression	0	0.887	0.422	0.813	0.887	0.848	0.494	0.817	0.879
	1	0.578	0.113	0.712	0.578	0.638	0.494	0.817	0.739
	Weighted Avg	0.786	0.321	0.780	0.786	0.779	0.494	0.817	0.833

As seen in Table 4, logistic regression has the highest *TP* rate (0.786), which means that 78.6% of the data defined are classified as a given class. Naive Bayes has the lowest *FP* rate (31.1%), which means that 31.1% of data is falsely classified as a given class. According to the *f*-measure, while logistics regression has the highest value (*f* = 0.779), random forest has the lowest value (*f* = 0.740). Across the J48 and multilayer perceptron algorithms, precision and *f*-measure exhibit the same pattern (≈ 0.77). To summarize, logistic regression has the highest values for *TP* rate, precision, recall, *f*-measure, *MCC*, *ROC* area, and *PRC* area when compared to the other algorithms.

In order to illustrate the classification successes of algorithms, *ROC* curves are drawn and given in Figure 1. The values of *ROC* areas are over 0.750 for all algorithms; this result shows that good classifications have been achieved. However, the *ROC* area in logistic regression is higher compared to the other algorithms. The logistic regression produces a significantly higher *ROC* area (0.833).



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

Figure 1. The *ROC* areas of (a) random forest, (b) J48, (c) naive Bayes, (d) multilayer perceptron, and (e) logistic regression.

In addition to comparing different performance measures, the confusion matrix shows the overall correct-classifications ratio for each algorithm (see Table 5). Correct-classifications ratios are greater than 74% for all algorithms. In all algorithms, correct-classifications ratios for students with mathematics scores less than the country average are between 85% and 91%, while correct-classifications ratios for students with mathematics score higher than the country average are between 52% and 61%. The overall correct-classifications ratios for the multilayer perceptron and J48 algorithms are 78.1% and 78.3%, respectively. Random forest algorithm has the lowest correct-classifications ratio (74.8%). Logistic regression has the highest correct-classifications ratio (78.6%). According to both Tables 4 and 5, one can conclude the use of logistic regression algorithm to be logical for classifying students according to their mathematics achievement.

Table 5
Confusion Matrix for Classification Algorithms

Classified as	Random Forest		J48		Naive Bayes		Multilayer Perceptron		Logistic Regression	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
a = 0	3,609	600	3,828	381	3,558	651	3,757	452	3,732	477
b = 1	976	1,065	975	1,066	789	1,252	917	1,124	861	1,180

Up to now, classification results have been compared using relatively different performance measures. In order to test whether these algorithms statistically differ in terms of their performance measures, Friedman's two-way analysis of variance, a non-parametric alternative of mixed-effects ANOVA (Reinard, 2006), is performed. For this purpose, in addition to the kappa, *MAE*, and *RMSE* values for each algorithm, the weighted average of all performance measures as given in Table 4 are used. The test statistic is found to be and the null hypothesis, which states that no difference exists between algorithms, is rejected at a 95% confidence level (associated *p*-value = .011). This means the performance measures significantly differ across the five algorithms. To determine where they significantly differ, the procedure continues with post-hoc pairwise comparisons tests. Comparisons that are found statistically significant at the 5% and 10% levels are shown in Table 6. As seen, the J48 algorithm statistically differs from the random forest at a 90% confidence level in terms of all

Table 6
Pairwise Comparisons

Algorithms	Test Statistics
Random Forest vs. J48	-1.273***
Random Forest vs. Naive Bayes	-0.455
Random Forest vs. Multilayer Perceptron	-1.00
Random Forest vs. Logistic Regression	-2.273**
J48 vs. Naive Bayes	0.818
J48 vs. Multilayer Perceptron	0.273
J48 vs. Logistic Regression	-1.00
Naive Bayes vs. Multilayer Perceptron	-0.545
Naive Bayes vs. Logistic Regression	-1.818**
Logistic Regression vs. Multilayer Perceptron	-1.273***

** *p*- < .05; *** *p* < .1.

performance measures. Furthermore, logistic regression is concluded to significantly differ from the random forest and naive Bayes algorithms at a 5% significance level, and from the multi-layer perceptron algorithm at a 10% significance level.

The last step of analysis is finding the subset of factors that produce the best classification and prediction performance. This procedure is done by sorting factors with respect to their discriminative power. WEKA includes several attitude-selection methods. The methods of correlation-ranking filter, gain ratios, and info gain have been used to find the factors' order of importance. Table 7 gives the obtained results. As seen, both selection methods produce the same ordered results for the seven factors. The top seven factors can be written as follows: SC (students confident in math), HER (home educational resources), PHEL (parents' highest education level), SLL (students like learning math), HSS (number of home study supports), SE (students engaged in math) and SVL (students value math). Among them, student confidence is found to be the most effective factor on mathematics achievement according to the three attitude-selection methods.

Table 7
Factor Ranking

Correlation Ranking Filter		Gain Ratio		Info Gain	
Rank	Name	Rank	Name	Rank	Name
1	SC	1	SC	1	SC
2	HER	2	HER	2	HER
3	PHEL	3	PHEL	3	PHEL
4	SLL	4	SLL	4	SLL
5	HSS	5	HSS	5	HSS
6	SE	6	SE	6	SE
7	SVL	7	SVL	7	SVL
8	SB	8	AGE	8	AGE
9	AGE	9	SB	9	SB
10	WTSMH	10	WTSMH	10	WTSMH
11	SEX	11	SEX	11	SEX

Conclusion and Discussion

Using data mining techniques in the field of education provides educators and education planners with a better understanding of huge data sets that include hidden useful patterns. Due to the fact that traditional methods have some limitations, such as in the cases of nonlinearity, non-homogeneity and non-normality, data mining methods have attracted much of researchers' attention. In recent years, several studies falling under the EDM concept have been done in order to investigate students' academic performances at the national level. TIMSS is an international assessment study that provides information about the effects of policy and practice in each participating country's education system (Mullis et al., 2012). This study aims to fill the gap in the current literature by applying different supervised algorithms to the TIMSS 2011, the last released version of this dataset. This dataset contains information on a total of 6,250 eighth-grade Turkish students. Mathematics achievement is a binary variable that receives a value of 0 or 1 and has been taken as the dependent variable. Eleven factors (one continuous and ten categorical) have been taken as the independent variables that can potentially affect mathematics achievement. Three research questions have been taken into account.

The first research question is "Which algorithm has the best classification performance based on model performance measures?" To answer this question, algorithms that are widely used in EDM literature are first chosen. Two decision-tree algorithms (random forest and J48), a Bayesian network algorithm (naive Bayes), an artificial neural-networks algorithm (multilayer perceptron), and logistic regression are performed on the data. According to several important classification-performance measures, the best performing algorithm is found as logistic regression. Additionally, the test results from Friedman's two-way analysis of variance show that logistic regression statistically differs from the other algorithms in terms all performance measures used in this study.

The second research question of "Which factors are found significant on mathematics achievement and what is their order of importance?" is answered using WEKA attitude-selection methods (i.e., correlation ranking filter, gain ratios, and info gain). All selection methods produce the most important factor to be students confidence. Home educational resources, parents' highest education level, students like learning, number of home study supports, student engagement, and students value learning appreciation have been found as important factors on mathematic achievement.

In order to deal with the third research question that asks what measures should be taken to improve the mathematics achievement of Turkish eighth-graders, one should focus on the factors found to be important on achievement. This study's findings are consistent with the literature. For example, students confidence has been referred to as an important predictor of academic achievement in the studies of Liu and Meng (2010), Hammouri (2010), and Aşkın and Gökalp (2013). The educational system should clearly be focused on enhancing students' confidence, and students should be more motivated in order to be more successful. Additionally, the association between parents' highest education level and academic achievement has been found significant in many studies (Berberoğlu, Çelebi, Özdemir, Uysal, & Yayan 2003; Topçu et al., 2016; Wößmann, 2005). This is a general idea, not only for Turkish eighth-grade students, but also for all students in the education system. Higher parental educational levels positively correlate with students' academic achievement. Topçu et al. (2016) suggested evening classes or summer institutions for students' lesser educated parents in order to increase their knowledge and awareness. Also, educational resources are another important indicator of success, and students who can reach educational resources easily are generally included in successful groups. Yıldırım and Demir (2014) included the factors of students value learning and students engagement in their multilevel models, reporting that these factors correlate with students' TIMSS scores. According to Fredricks, Blumenfeld, and Paris (2004), motivated students are more consistent when encountering difficulties and prefer challenges; this is why the factor of students' value learning is highly correlated with mathematics achievement.

In this study, different classification performances have been compared for the dataset of Turkish eighth-grade students' mathematics scores from the TIMSS 2011, with three research questions being answered. Further research and applications can focus on the results of various countries that have participated in the TIMSS. The TIMSS study also not only deals with mathematics achievement, but also assesses students' science and reading skills. Therefore, studies can be done on different educational areas for the purpose of finding factors that influence achievement in the related courses. Additionally, the algorithms given in this study can be performed in different national/international assessment studies, such as PISA and PIRLS, in order to understand the superiority of these data-mining algorithms as alternatives to standard statistical procedures.

References

Abad, F., & Lopez, A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement, 27*(1), 39–55. http://dx.doi.org/10.1080/09243453.2016.1235591

Alivernini, F. (2013). An exploration of the gap between highest and lowest ability readers across 20 countries. *Educational Studies*, *39*(4), 399–417. http://dx.doi.org/10.1080/03055698.2013.767187

Aşkın, Ö. E., & Gökalp, F. (2013). Comparing the predictive and classification performances of logistic regression and neural networks: A case study on TIMSS 2011. *Procedia-Social Behavioral Science, 106*, 667–676. http://dx.doi.org/10.1016/j.sbspro.2013.12.076

Baker, R. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International encyclopedia of education* (3rd ed., vol. 7, pp. 112–118). Oxford, UK: Elsevier.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Berberoğlu, G., Çelebi, Ö., Özdemir, E., Uysal, E., & Yayan, B. (2003). Factors affecting achievement level of Turkish students in the Third International Mathematics and Science Study. *Educational Sciences and Practice*, *2*(3), 3–14.

Bilen, Ö., Hotaman, D., Aşkın, Ö. E., & Büyüklü, A. H. (2014). Analyzing the school performances in terms of LYS successes through using data mining techniques: Istanbul sample, 2011. *Education and Science, 39*(172), 78–94.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. http://dx.doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random forest. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324

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

Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44–48. http://dx.doi.org/10.5120/7228-0076

Davis, J., & Goadrich, M. (2006, June). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). New York, NY: ACM.

Dejaeger, K., Goethals, F., Giangreco, A., Mola, L., & Baesens, B. (2012). Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, *218*(2), 548–562. http://dx.doi.org/10.1016/j.ejor.2011.11.022

Donner, A., & Klar, N. (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology*, *49*(9), 1053–1058. http://dx.doi.org/10.1016/0895-4356(96)00057-1

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. http://dx.doi.org/10.3102/00346543074001059

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*(3), 241–257. http://dx.doi.org/10.1080/0013188042000277313

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

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, *29*(1), 90–102. http://dx.doi.org/10.1016/j.chb.2012.07.020

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

İdil, F., Narlı, S., & Aksoy, E. (2016). Using data mining techniques examination of the middle school students' attitude towards mathematics in the context of some variables. *International Journal of Education in Mathematics, Science and Technology*, *4*(3), 210–228. http://dx.doi.org/10.18404/ijemst.02496

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338–345). Burlington, MA: Morgan Kaufmann Publishers.

Kılıç, S., & Aşkın, Ö. E. (2013). Parental influence on students' mathematics achievement: The comparative study of Turkey and best performer countries in TIMSS 2011. *Procedia-Social Behavioral Sciences*, *106*, 2000–2007. http://dx.doi.org/10.1016/j.sbspro.2013.12.228

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2010). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411–426. http://dx.doi.org/10.1080/08839510490442058

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*(6), 699–712. http://dx.doi.org/10.1080/01443410.2010.501102

Mohamad, S., & Tasir, Z. (2013). Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, *97*, 320–324. http://dx.doi.org/10.1016/j.sbspro.2013.10.240

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: Boston College.

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995–2003). *Studies in Educational Evaluation*, *34*(2), 56–72. http://dx.doi.org/10.1016/j.stueduc.2008.04.002

Öz, E., Kurt, S., Asyalı, M., Kaya, H., & Yücel, Y. (2016). Feature based quality assessment of DNA sequencing chromatograms. *Applied Soft Computing*, *41*, 420–427. http://dx.doi.org/10.1016/j.asoc.2016.01.025

Pena-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert System with Applications*, *41*(4), 1432–1462. http://dx.doi.org/10.1016/j.eswa.2013.08.042

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Burlington, MA: Morgan Kaufmann Publishers.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science Issues*, *7*(1), 10–18. http://dx.doi.org/10.1.1.403.8058

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. *International Journal of Computer Applications*, *63*(8), 35–39. http://dx.doi.org/10.5120/10489-5242

Razi, M., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, *29*(1), 65–74. http://dx.doi.org/10.1016/j.eswa.2005.01.006

Reinard, J. (2006). *Communication research statistics*. London, UK: Sage.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. http://dx.doi.org/10.1016/j.eswa.2006.04.005

Schreiber, J. (2002). Scoring above the international average: A logistic regression model of the TIMSS advanced mathematics exam. *Multiple Linear Regression Viewpoints*, *28*(1), 22–30.

Shahiri, A., Husain, W., & Rashid, N. (2015). A review on predicting students' performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422. http://dx.doi.org/10.1016/j.procs.2015.12.157

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, *35*(3), 38–54. http://dx.doi.org/10.1111/emip.12088

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, *36*(2), 111–147. http://dx.doi.org/10.2307/2984809

Sugumaran, V., Muralidharan, V., & Ramachandran, K. I. (2007). Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing*, *21*(2), 930–942. http://dx.doi.org/10.1016/j.ymssp.2006.05.004

Topçu, M., Erbilgin, E., & Arıkan, S. (2016). Factors predicting Turkish and Korean students' science and mathematics achievement in TIMSS 2011. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(7), 1711–1737. http://dx.doi.org/10.12973/eurasia.2016.1530a

Turanoğlu-Bekar, E., Ulutagay, G., & Kantarcı-Savaş, 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*, *2016*(2), 13–28. http://dx.doi.org/10.5899/2016/ojids-00002

Vialardi, C., Chue, J., Peche, J., Alvarado, G., Vinatea, B., Estrella, J., … Ortigosa, A. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User-Adapted Interaction*, *21*(1), 217–248. http://dx.doi.org/10.1007/s11257-011-9098-4

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*(1), 79–82. http://dx.doi.org/10.3354/cr030079

Wößmann, L. (2005). Educational production in East Asia: The impact of family background and schooling policies on student performance. *German Economic Review*, *6*(3), 331–353. http://dx.doi.org/10.1111/j.1468-0475.2005.00136.x

Yıldırım, Ö., & Demir, S. B. (2014). The examınatıon of teacher and student effectiveness at TIMSS 2011 Science and math scores using multi level models. *Pakistan Journal of Statistics, 30*(6), 1211–1218.