IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 6, 2016, accepted May 10, 2016, date of publication May 13, 2016, date of current version June 3, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2568756

Predicting Instructor Performance Using Data Mining Techniques in Higher Education

MUSTAFA AGAOGLU
Department of Computer Engineering, Marmara University, Istanbul 34722, Turkey (agaoglu@marmara.edu.tr).

ABSTRACT Data mining applications are becoming a more common tool in understanding and solving educational and administrative problems in higher education. In general, research in educational mining focuses on modeling student's performance instead of instructors' performance. One of the common tools to evaluate instructors' performance is the course evaluation questionnaire to evaluate based on students' perception. In this paper, four different classification techniques—decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis—are used to build classifier models. Their performances are compared over a data set composed of responses of students to a real course evaluation questionnaire using accuracy, precision, recall, and specificity performance metrics. Although all the classifier models show comparably high classification performances, C5.0 classifier is the best with respect to accuracy, precision, and specificity. In addition, an analysis of the variable importance for each classifier model is done. Accordingly, it is shown that many of the questions in the course evaluation questionnaire appear to be irrelevant. Furthermore, the analysis shows that the instructors' success based on the students' perception mainly depends on the interest of the students in the course. The findings of this paper indicate the effectiveness and expressiveness of data mining models in course evaluation and higher education mining. Moreover, these findings may be used to improve the measurement instruments.

INDEX TERMS Artificial neural networks, classification algorithms, decision trees, linear discriminant analysis, performance evaluation, support vector machines.

I. INTRODUCTION

Today, one of the biggest challenges of higher education institutions is the proliferation of data and how to use them to improve quality of academic programs and services and the managerial decisions [1]–[3]. A variety of ''formal and informal'' procedures based on ''qualitative and quantitative'' methods is used by higher education institutions to solve problems, which keep them away from achieving their quality objectives [1], [2]. However, methods used in higher education for quality purposes are mainly based on predefined queries and charts to analyze the data. In addition, these methods lack the ability to reveal useful hidden information [1].

Hidden information in large datasets is best analyzed with data mining techniques. Data mining (sometimes called knowledge discovery) is the process of discovering ''hidden messages,'' patterns and knowledge within large amounts of data and process of making predictions for outcomes or behaviors [4]. Data mining can be best defined as the automated process of extracting useful knowledge and information including patterns, associations, changes, trends, anomalies, and significant structures that are unknown from large or complex datasets [5].

Lately, there is an increased popularity of using data mining techniques in higher education, and because of its potentials to educational institutes such as better allocating resources [6], predicting student performance [7], academic planning and intervention transfer prediction [6], improving the effectiveness of alumni development [4]; a new field called educational data mining has emerged [8], [9]. Educational data mining (EDM) is concerned with developing methods for exploring data from educational settings with the purpose of providing quality education to students [10]. With EDM, additional insights can be gained from educational entities such as students, lecturers, staff, alumni, and managerial behavior [2]. These can be then used to allocate resources and staff more effectively, make better decisions on educational activities to improve students' success, increase students' learning outcome, increase student's retention rate, decrease students' drop-out rate, and reduce the cost of system processes [2], [3].

One of the common problems in higher education is the evaluation of instructors' performances in a course. The most widely applied tool to evaluate the instructors' performance in a course is through surveying students' responses about

the course and its instructor through a questionnaire. Since 1920's, when student evaluations of instructor performance were first introduced to higher education systems, there has been an ongoing debate on the reliability and validity of these evaluations [11], [12]. The concern about student evaluations is based mainly on 1) students not having enough experience and maturity to evaluate the course and the instructor, 2) students' evaluations being affected by popularity of the course and/or instructor, grades given by the instructor and course being compulsory or elective [13], [14].

However, students are the only source of information about the learning environment, and they are the only ones who can rate the quality, the effectiveness, and the satisfaction of course content, method of instruction, textbook, and homework [13], [15]. Student evaluations are used primarily to improve course and teaching quality as well as a part of the evaluation process for staff appraisal systems [16], [17]. Also despite the discussions, several past studies found that student evaluation of teaching offers a reliable and valid assessment of instructors [18]–[21]. Since the majority of universities use student evaluations as the measure of effectiveness [14], [22], [23], students' opinions about the courses and instructors are collected and stored in databases waiting to be discovered and used for managerial purposes.

Yet, studies on student evaluations are mainly concerned on psychometric properties of the instruments used, effect of these assessments on instructor effectiveness, and usefulness and validity for both formative and summative purposes where statistical techniques are employed [14], [23]–[25].

On the other hand, even though there is an increase in EDM studies, they mainly focus on models for exploring learning environments, web-based educational systems, improving student performance, and restructuring curricula [9], [10], [26]. Furthermore, they do not include student evaluations except a few [1], [27], [28].

The aim of this research is to show the potential of EDM in enlightening the criteria or measures of effective instructor performance as perceived by the students. In this study, four classification techniques –decision tree algorithms, support vector machines (SVM), artificial neural networks (ANN), and discriminant analysis (DA)– are chosen to build classifier models on a dataset composed of the responses of students to a course evaluation questionnaire and the performances of these models are compared.

Classification techniques are not widely used in EDM literature until 2009 [7], [29], however there is an increase in their application within the last six years [26]. Nonetheless, researchers prefer to apply a single technique in their studies on student evaluations like those mentioned above. This study not only employs different classification techniques to student evaluation data but also make their performance comparisons with respect to several metrics by indicating outstanding method(s) within classification techniques when applied in this field. As a novelty, boosting is also used for instructor performance evaluation.

The rest of this paper is organized as follows: Section II gives a review of the classification models and techniques used in this study. Section III gives some insights to the structure of data and data collection process. Section IV presents the results and discussions, and Section V concludes the study.

II. CLASSIFICATION MODELS

Classification is one of the most common application domains of data mining. The main task in classification is assigning a class label among a set of possible class values to an unseen instance composed of a set of variables. It is done by using a classifier model, which is built by applying a learning algorithm on a training set composed of past instances having the same variable set as the unseen instance. However, the class label of each instance in the training set is clearly known before training. After learning phase, the classification performance of the classifier model built is evaluated on an independent test set before used.

In classification, there are many different methods and algorithms possible to use for building a classifier model. Some of the most popular ones can be counted as decision tree algorithms, support vector machines (SVM), artificial neural networks (ANN), discriminant analysis (DA), logistic regression, Bayesian belief networks, and rule based systems. In this study, the first four of these are used.

A decision tree algorithm aims to recursively split the observations into mutually exclusive subgroups until there is no further split that makes a difference in terms of statistical or impurity measures. Among the impurity measures that are used to find the homogeneity of instances in a node of the tree, Information Gain, Gain Ratio, and Gini Index are the most well-known ones. Usually, Information Gain is used in Iterative Dichotomiser (ID3), Gain Ratio in C4.5 and C5.0 [30] (the successors of ID3) whereas Gini Index is used in Classification and Regression Trees (CART [31]). Fig. 1 shows a sample representation for a decision tree.

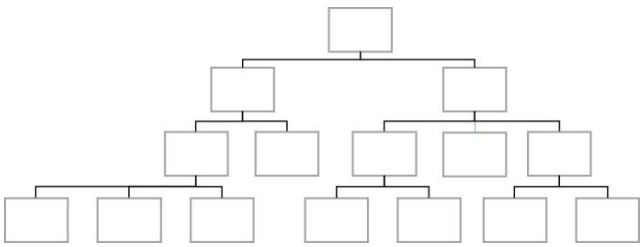


FIGURE 1. Decision tree diagram.

In the following equation sets, D stands for database or whole dataset, and A denotes an attribute/variable of the entity. $Info(D)$ is the amount of information needed to identify the class label of a tuple in D for the whole dataset whereas $Info_A(D)$ is the information for an attribute that is to be calculated. The probability of an arbitrary tuple in D belongs to class C_i is denoted by p_i values starting from first one to the last ($i = 1 \ldots m$). When we are supposed to partition the

tuples in D on some attribute A, which has v distinct values based on the training dataset, we can split D into v subsets, given by D_j. Information Gain is then defined, given in (3), as the difference between the information of the whole dataset and the new case obtained after partitioning on A.

Equations (1), (2), and (3) are used for the calculations of Information Gain, which is the impurity measure of ID3:

$$Info\,(D) = -\sum_{i=1}^{m} p_i \log_2 (p_i) \tag{1}$$

$$Info_A\,(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2}$$

$$Gain\,(A) = Info\,(D) - Info_A(D) \tag{3}$$

In C4.5 and C5.0 algorithms, there are some improvements in terms of handling the bias toward tests with many outcomes, performance, and pruning. These algorithms use Gain Ratio as an extension to ID3 calculations given in (4) and (5), which has a kind of normalization to Information Gain using *SplitInfo* values.

$$SplitInfo_A\,(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \tag{4}$$

$$GainRatio\,(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{5}$$

CART algorithm uses Gini index as an impurity measure and some calculations are given in (6) and (7). Using the same notations as in the calculations of Information Gain, the impurity of D is measured in (6) whereas the reduction in impurity that would be obtained by a split on attribute A is given in (7).

$$Gini\,(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{6}$$

$$\Delta Gini\,(A) = Gini\,(D) - Gini_A(D) \tag{7}$$

As a novel approach, boosting, an ensemble learning method, which builds a strong classifier using a set of iteratively learning weak classifiers, is used with C5.0 classifiers as base classifiers to further increase the performance. Idea behind boosting is building a strong classifier using a set of iteratively learning weak classifiers by adding a new weak classifier and weighting the data to focus on misclassified entities for the new weak learners [32], [33].

Unlike decision tree algorithms, SVM try to find a hyperplane to separate the classes while minimizing the classification error and maximizing the margins. SVM is a good classification and regression technique proposed by Vapnik at AT&T Bell Laboratories [34]. Fig. 2 shows a sample representation for SVM.

ANN, another technique used in this study, is a group of nonlinear, statistical modeling techniques, which is inspired and derived from human brain. Similar to human brain, ANNs are mostly shown as a system of interconnected neurons that exchange information between each other. The connections between the neurons have varying weights, and these can be tuned based on the given information (i.e. training dataset) which makes ANN capable of learning.

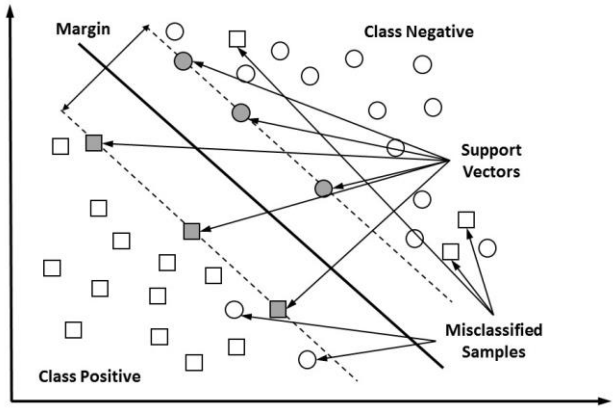


FIGURE 2. Support vector machines diagram.

ANNs are generally defined by three parameters: 1) The interconnection pattern and weights between the nodes and different layers of neurons, 2) The learning process for updating the weights, and 3) The activation function that converts a neuron's weighted input to its output. A sample neural network with one hidden layer is given in Fig. 3.

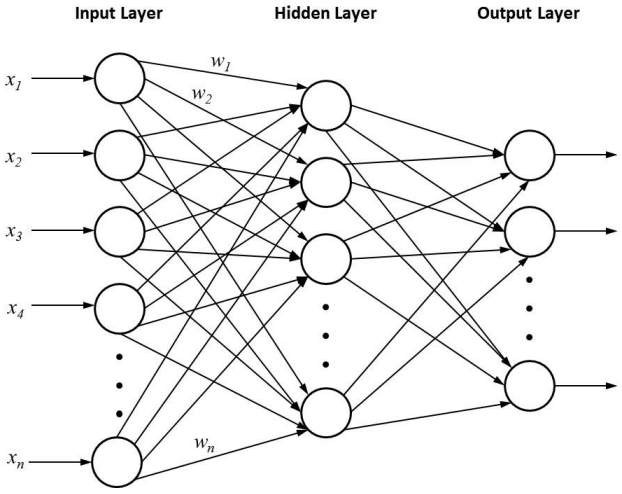


FIGURE 3. Artificial neural network diagram.

ANN needs long training time to construct a well-suited model and it is hard to interpret ANN because of its nodes and hidden layer structure. But fortunately, it can tolerate noisy data and can be used even if when there is no relationship between variables and classes. Because of the fact that ANN can be used in any complex pattern modeling, it strongly fits to any classification problem.

Discriminant analysis (DA), the last technique used in this study, is a statistical data mining technique used to predict a categorical response variable [5]. It is one of the classification methods, where observations are assigned to one of the predefined groups based on the knowledge of the variables with the purpose of profiling or differentiation [35]–[37]. DA has the underlying assumptions of multivariate normality

and all groups having equal covariance structures [38], [39]. However, DA has been found to be very robust to deviations from these assumptions in practice [35].

Two complementary approaches to DA are Fisher's and Mahalanobis' [37]. Fisher's approach calculates a linear combination of the independent variables such that the ratio of the across group variation to the within group variation in discriminant score is maximized. Mahalanobis' approach calculates the covariance adjusted distance from a point to each group centroid and assigning it to the closest group. When the class variable has two categories, the two approaches give equivalent results [37].

There are some performance measures to evaluate classification models in terms of the correctness of the classification decisions of the model. Assuming a binary classification task as in this study, the class variables values may be assumed as Positive (P) and Negative (N). Actual positives (P) that are correctly labeled as positives by the classifier are named as true positives (TP) whereas actual positives incorrectly labeled as negatives by the classifier are considered as false negatives (FN). In a similar fashion, actual negatives (N) that are correctly labeled as negatives are taken as true negatives (TN) whereas actual negatives incorrectly labeled as positives are considered as false negatives (FP). These terms are given in the confusion matrix of Table. 1.

TABLE 1. Confusion matrix.

		Predicted Class		
		Positive	Negative	Total
Actual Class	Positive	TP	FN	P
	Negative	FP	TN	N
	Total	P'	N'	P + N

The calculations of performance measures such as accuracy (recognition rate), precision, recall (sensitivity or true positive rate), and specificity (true negative rate) are given in (8), (9), (10), and (11). Accuracy measures the rate of total correct predictions to all predictions. Precision measures the correctness rate of the class predictions done as positive by the classifier whereas recall measures the rate of positives correctly predicted as positive by the classifier. Likewise, specificity measures the rate of negatives correctly predicted as negative by the classifier.

$$Accuracy = \frac{TP + TN}{P + N} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{P} \tag{10}$$

$$Specificity = \frac{TN}{N} \tag{11}$$

III. DATA COLLECTION AND VARIABLES

Data is collected from one of the randomly selected departments of Marmara University, Istanbul, Turkey. A total of 2850 evaluation scores are obtained. In the data mining process, randomly selected 70% of these, 1995 observations, are used to train the classifier models. The remaining 30%, 855 observations, are used as the test data.

Student evaluation data has 26 variables all except one, which is class label, are responses, measured on an interval scale, to questions in course and instructor performance evaluation forms. Response values of these questions are of the form $\{1, 2, 3, 4, 5\}$ where $1, 2, 3, 4, 5$ represents the answers "Never", "Rarely", "Sometimes", "Often", "Always" respectively for Q1 to Q4; and $1, 2, 3, 4, 5$ represents "Strongly disagree", "Disagree", "Neutral", "Agree", and "Strongly agree" respectively for Q5 to Q25. The last variable is dichotomous variable measured on a nominal scale in the form of $\{1, 2\}$ where 1 stands for "Not satisfactory" and 2 for "Satisfactory". Details of the variables used in this study are shown in Table. 2.

TABLE 2. Details of the variables.

Variable	Description	Possible values
Q1	How frequently did you attend this course?	{1, 2, 3, 4, 5}
Q2	How frequently did the instructor attend this course?	{1, 2, 3, 4, 5}
Q3	Did the instructor come to class on time?	{1, 2, 3, 4, 5}
Q4	Did the instructor use the class hours fully?	{1, 2, 3, 4, 5}
Q5	The instructor's knowledge of the subject was adequate.	{1, 2, 3, 4, 5}
Q6	The instructor came to class well prepared.	{1, 2, 3, 4, 5}
Q7	Lectures were clear and well structured.	{1, 2, 3, 4, 5}
Q8	The syllabus was followed and fully covered.	{1, 2, 3, 4, 5}
Q9	There was a good balance between theory and application.	{1, 2, 3, 4, 5}
Q10	The instructor was effective in presenting the material in lectures and discussions.	{1, 2, 3, 4, 5}
Q11	Subjects were not well organized and did not follow each other.	{1, 2, 3, 4, 5}
Q12	The instructor answered questions satisfactorily.	{1, 2, 3, 4, 5}
Q13	The instructor was enthusiastic about this course.	{1, 2, 3, 4, 5}
Q14	The instructor motivated students to ask questions, to get involved in discussions.	{1, 2, 3, 4, 5}
Q15	The instructor was approachable, nice, and easy to communicate with.	{1, 2, 3, 4, 5}
Q16	The instructor was available to give help outside the class (via mail or in person).	{1, 2, 3, 4, 5}
Q17	The course materials (textbooks, handouts, etc.) were satisfactory.	{1, 2, 3, 4, 5}
Q18	The assignments (case studies, homework, projects, presentations, etc.) were relevant and helpful.	{1, 2, 3, 4, 5}
Q19	This was an easy AA course.	{1, 2, 3, 4, 5}
Q20	The assignments/exams increased the ability to think creatively.	{1, 2, 3, 4, 5}
Q21	The assignments/exams adequately tested the knowledge taught.	{1, 2, 3, 4, 5}
Q22	I had to spend a lot of time to get prepared for assignments/exams.	{1, 2, 3, 4, 5}
Q23	This course was helpful in increasing my knowledge and interest in the subject matter.	{1, 2, 3, 4, 5}
Q24	I learned a lot in this course.	{1, 2, 3, 4, 5}
Q25	The instructor's grading (assignments/exams) was fair.	{1, 2, 3, 4, 5}
Q26	The overall performance of the instructor was satisfactory.	{1, 2}

Variable Q26 is the class variable that is to be predicted. The value ''satisfactory'' for this variable is taken as positive class label whereas ''not satisfactory'' is assumed to be negative.

IV. RESULTS AND DISCUSSIONS

In this study, seven classification models are generated: two using decision tree algorithms (C5.0, and CART), one using SVM, three using ANNs, and one using DA. The performances of these models are evaluated on the test data in terms of accuracy, precision, recall, and specificity.

In the build settings of C5.0 classifier, Gain Ratio is used as impurity measure and splits are done with multiway splits. As the stopping rule, ''minimum instances per child'' is set as two. In addition, boosting is applied with 30 trials on C5.0 algorithm. The confusion matrix showing the distribution of predictions of C5.0 classifier is given in Table. 3. We have 457 TP, and 332 TN instances whereas 27 FP, and 39 FN instances as the classification result of C5.0 classifier.

TABLE 3. Confusion matrix of C5.0.

	Satisfactory	Not satisfactory	Total
Satisfactory	457	39	496
Not satisfactory	27	332	359
Total	484	371	855

Rows represent actual, columns represent predicted group memberships

In the build settings of CART classifier, Gini Index is used as impurity measure and splitting is done with binary splits. As the stopping rule, minimum change in impurity is set as 0.0001. The confusion matrix showing the distribution of predictions of CART classifier is given in Table. 4. We have 462 TP, and 307 TN instances whereas 52 FP, and 34 FN instances as the classification result of CART classifier.

TABLE 4. Confusion matrix of CART.

	Satisfactory	Not satisfactory	Total
Satisfactory	462	34	496
Not satisfactory	52	307	359
Total	514	341	855

Rows represent actual, columns represent predicted group memberships

In the build settings of support vector machines (SVM) classifier, a fifth degree polynomial function is used as the kernel function. The confusion matrix showing the distribution of predictions of SVM classifier is given in Table. 5. We have 461 TP, and 320 TN instances whereas 39 FP, and 35 FN instances as the classification result of SVM classifier.

TABLE 5. Confusion matrix of SVM.

	Satisfactory	Not satisfactory	Total
Satisfactory	461	35	496
Not satisfactory	39	320	359
Total	500	355	855

Rows represent actual, columns represent predicted group memberships

In the build settings of ANN classifiers, two different feed-forward backpropagation ANN methods implemented in IBM SPSS Modeler, Quick and Multiple are used. Two classifiers with unique settings are built using Quick method and another classifier is built using Multiple method. For the first Quick classifier (ANN-Q2H), a topology composed of two hidden layers with 20 and 15 nodes respectively is used; whereas for the second Quick classifier (ANN-Q3H), a topology composed of three hidden layers with 20, 15, and 10 nodes respectively is used. In Multiple method, multiple networks are trained in pseudo parallel approach. When the stopping criterion is met for all networks, the network with the highest accuracy is returned as the final model. The final classifier of Multiple method (ANN-M) is composed of two hidden layers with 67 and 5 nodes respectively is used. The confusion matrices showing the distribution of predictions of classifiers ANN-Q2H, ANN-Q3H, and ANN-M are given in Table. 6, 7, and 8, respectively. In the result of ANN-Q2H classifier, we have 449 TP, and 331 TN instances whereas 28 FP, and 47 FN instances. Similarly, there are 471 TP, and 305 TN instances whereas 54 FP, and 25 FN instances as the result of ANN-Q3H. Moreover, for the ANN-M classifier, we have 454 TP, and 320 TN instances whereas 39 FP, and 42 FN instances.

TABLE 6. Confusion matrix of ANN-Q2H.

	Satisfactory	Not satisfactory	Total
Satisfactory	449	47	496
Not satisfactory	28	331	359
Total	477	378	855

Rows represent actual, columns represent predicted group memberships

TABLE 7. Confusion matrix of ANN-Q3H.

	Satisfactory	Not satisfactory	Total
Satisfactory	471	25	496
Not satisfactory	54	305	359
Total	525	330	855

Rows represent actual, columns represent predicted group memberships

TABLE 8. Confusion matrix of ANN-M.

	Satisfactory	Not satisfactory	Total
Satisfactory	454	42	496
Not satisfactory	39	320	359
Total	493	362	855

Rows represent actual, columns represent predicted group memberships

Since DA is a statistical data mining technique; prior to the analysis, basic assumptions are tested. The assumptions related to formation of discriminant function, which are normality, linearity, and multicollinearity [40] are satisfied indicating that DA can be applied. However, assumption related to estimation of discriminant function, which is equal covariance matrices [40] is violated according to Box's M test ($Box's\ M = 548.62$, $F\ (55, 7216482) = 9.92$, $p = 0.00$). In DA, the most difficult assumption to meet is equal covariance matrices especially when the sample size is large,

but DA is a rather robust technique that can tolerate this assumption [36], [39], [40]. Therefore, we continue DA with build settings taken as: prior probabilities set all equal groups, method as stepwise, and DA approaches Fisher's and Mahalanobis'.

TABLE 9. Result of stepwise DA.

Step	Variable Entered	Min. D^2	Wilks' Lambda	F value	p value	df1	df2
1	Q6	2.521	0.629	1111.518	0.000	1	1883
2	Q24	3.487	0.551	768.242	0.000	2	1882
3	Q16	4.012	0.516	588.929	0.000	3	1881
4	Q25	4.310	0.498	474.186	0.000	4	1880
5	Q13	4.491	0.488	395.062	0.000	5	1879
6	Q5	4.604	0.481	337.366	0.000	6	1878
7	Q21	4.686	0.477	294.161	0.000	7	1877
8	Q8	4.748	0.474	260.675	0.000	8	1876
9	Q15	4.791	0.471	233.671	0.000	9	1875
10	Q23	4.824	0.470	211.647	0.000	10	1874

Table. 9 provides the stepwise discriminant analysis result, which shows all the significant variables that are included in the estimation of the discriminant function. Based on their Wilks' Lambda and minimum Mahalonobis D^2 values ten out of twenty-five variables are significant discriminators of positive and negative performance of instructors.

Canonical discriminant function is significant at $\alpha = 0.001$, indicating that the discriminant function can be used to classify positive and negative performance of instructors and canonical correlation which is 0.728 indicates the usefulness of discriminant function with its high score ($Wilk's\ \Lambda = 0.47$, $\chi 2\ (10) = 1419.46$, $p = 0.00$). Standardized and unstandardized coefficients of the discriminant function are displayed in Table. 10. The standardized coefficients reveal the relative importance of each variable.

TABLE 10. Canonical discriminant function coefficients.

Independent Variables	Standardized Coefficients	Unstandardized Coefficients
Q25	0.189	0.208
Q6	0.172	0.211
Q5	0.163	0.194
Q16	0.161	0.175
Q24	0.161	0.189
Q13	0.157	0.190
Q8	0.127	0.155
Q21	0.126	0.156
Q23	0.116	0.139
Q15	0.115	0.133
Constant		-6.708

Then the unstandardized coefficients are used to determine the discriminant scores, which are used in classification. To determine how well the discriminant function can predict group memberships, as we do with the other techniques a confusion matrix is formed, which is given in Table. 11. We have 445 TP, and 264 TN instances whereas 28 FP, and 46 FN instances as the result of DA classifier. In DA, missing values in the dataset are solved by case wise deletion therefore the total number of instances is less than 855, which is 783.

TABLE 11. Confusion matrix of DA.

	Satisfactory	Not satisfactory	Total
Satisfactory	445	46	491
Not satisfactory	28	264	292
Total	473	310	783

Rows represent actual, columns represent predicted group memberships

In this part, a performance comparison of all the classifiers applied is done using evaluation measures. As can be seen from Table. 12, classifiers give similar results on the test dataset. When we compare the model performances, we see that all the methods we used in this study are effective in classifying "satisfactory" and "not satisfactory" instructor performances.

TABLE 12. Performances of classifiers.

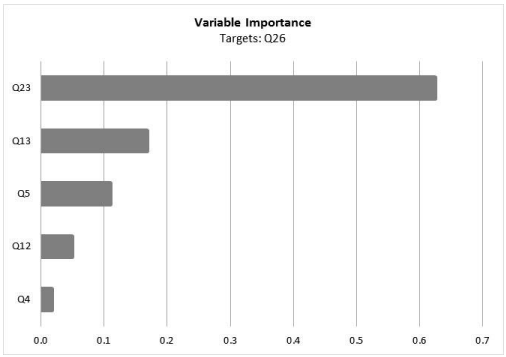
Model	Accuracy	Precision	Recall	Specificity
C5.0	**92.3 %**	**94.4 %**	92.1 %	**92.5 %**
CART	89.9 %	89.9 %	93.1 %	85.5 %
SVM	91.3%	92.2%	92.9%	89.1%
ANN-Q2H	91.2 %	94.1 %	90.5 %	92.2 %
ANN-Q3H	90.8 %	89.7 %	**95.0 %**	85.0 %
ANN-M	90.5 %	92.1 %	91.5 %	89.1 %
DA	90.5 %	94.1 %	90.6 %	90.4 %

Accuracy values, which assess the effectiveness of the models, are all at least approximately 90%. C5.0 classifier is the best in performance according to accuracy followed by SVM, and CART is the worst. Precision, which assesses the predictive power, again indicated C5.0 as the best classifier; however, ANN-Q2H and DA also show high predictive power. On the other hand, recall values, which indicate the sensitivity and true positive rate (TPR) of the models, differ among classifiers. According to recall, ANN-Q3H is the best classifier in performance and the CART is the second best. In addition, according to specificity, which is true negative rate (TNR), C5.0 is again is the best and ANN-Q2H is the second best. Moreover, ANN-Q3H is the worst in performance according to precision and specificity whereas ANN-Q2H is the worst in recall measures.

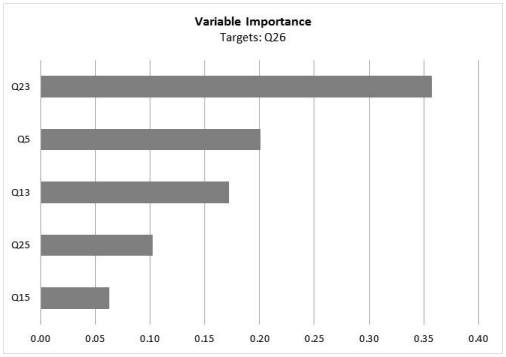
As a result, we can say even though all the classifiers are similarly good, from the two decision tree classifiers, C5.0, and from the three artificial neural network classifiers, the ANN-Q2H is comparatively better. Ultimately, C5.0 can be considered as the outstanding classifier among all according to the given performance measures.

Variable importance graphs of the first six classifiers and the standardized canonical discriminant function coefficients of DA are given in Fig. 4.

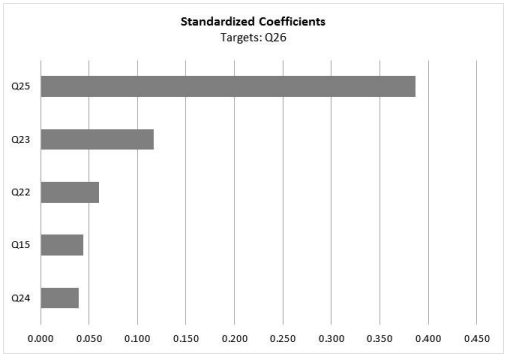
In C5.0 classifier, the top three important variables are Q23 "this course was helpful in increasing my knowledge and interest in the subject matter," Q13 "the instructor was enthusiastic about this course," and Q5 "the instructor's knowledge of the subject was adequate," which can be seen in Fig. 4(a).



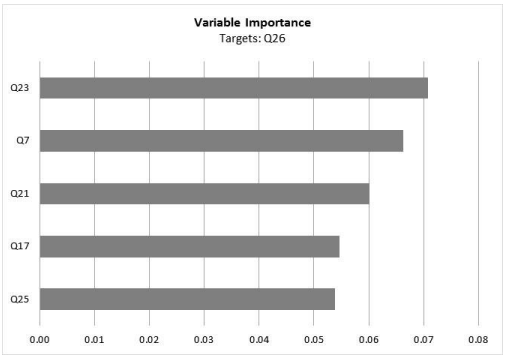
Importance of variables in C5.0
(a)



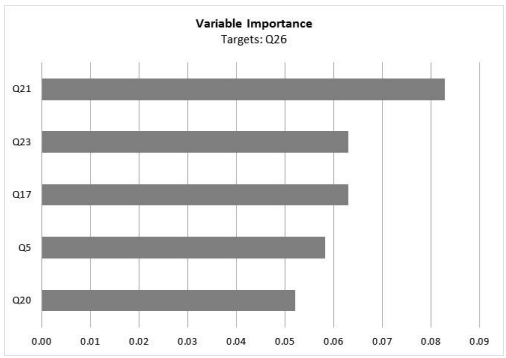
Importance of variables in CART
(b)



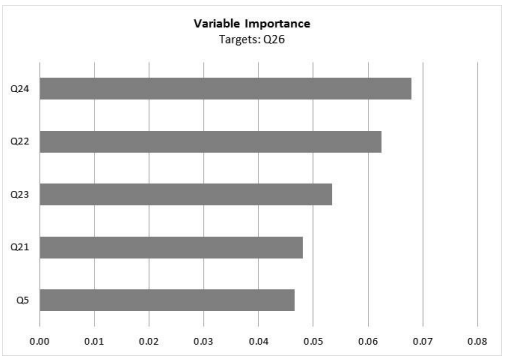
Importance of variables in SVM
(c)



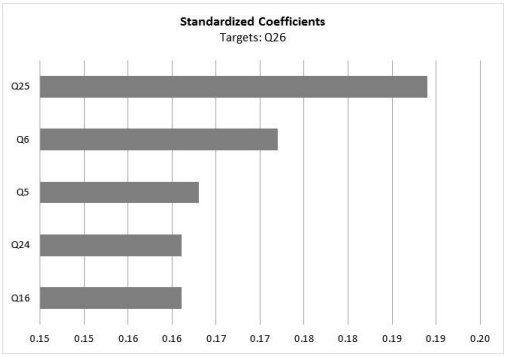
Importance of variables in ANN-Q2H
(d)



Importance of variables in ANN-Q3H
(e)



Importance of variables in ANN-M
(f)



Canonical Discriminant Function Coefficients in DA
(g)

FIGURE 4. Comparison of variables between classifiers by importance values in C5.0, CART, SVM, ANN-Q2H, ANN-Q3H, and ANN-M and standardized canonical discriminant function coefficients in DA. (a) Importance of variables in C5.0. (b) Importance of variables in CART. (c) Importance of variables in SVM. (d) Importance of variables in ANN-Q2H. (e) Importance of variables in ANN-Q3H. (f) Importance of variables in ANN-M. (g) Canonical discriminant function coefficients in DA.

In CART classifier, the most important three variables are again Q23, Q13, and Q5, which is shown in Fig. 4(b). However, this time Q5 is more important compared to Q13.

In SVM classifier, the most important three variables are Q25 ''the instructor's grading (assignments/exams) was fair,'' Q23 ''this course was helpful in increasing my knowledge and interest in the subject matter,'' and Q22 ''I had to spend a lot of time to get prepared for assignments/exams,'' which is given in Fig. 4(c).

In ANN-Q2H classifier, the top three important variables are Q23 ''this course was helpful in increasing my knowledge and interest in the subject matter,'' Q7 ''lectures were clear and well structured,'' and Q21 ''the assignments/exams adequately tested the knowledge taught,'' which is given in Fig. 4(d).

In ANN-Q3H classifier, the most important three variables are Q21 ''the assignments/exams adequately tested the knowledge taught,'' Q17 ''the course materials (textbooks, handouts, etc.) were satisfactory,'' and Q23 ''this course was helpful in increasing my knowledge and interest in the subject matter,'' which can be seen in Fig. 4(e).

In ANN-M classifier, the top three important variables are Q24 ''I learned a lot in this course,'' Q22 ''I had to spend a lot of time to get prepared for assignments/exams,'' and Q23 ''this course was helpful in increasing my knowledge and interest in the subject matter,'' which is shown in Fig. 4(f).

In DA classifier, the most important three variables are Q25 ''the instructor's grading (assignments/exams) was fair,'' Q6 ''the instructor came to class well prepared,'' and Q5 ''the instructor's knowledge of the subject was adequate,'' which is given in Fig. 4(g).

On the overall, Q23 ''This course was helpful in increasing my knowledge and interest in the subject matter,'' is the most important variable in differentiating positive and negative instructor performance since it was in top three variables in C5.0, CART, SVM, ANN-Q2H, ANN-Q3H, and ANN-M. Even though it is not in the top three list, Q23 is one of the significant discriminators in DA as well. Therefore, we can say that the interest area of students and the subject of the course are more important to students than instructors' behavior in evaluating the instructor performance.

Remarkably, eight of the variables either not distinguish the positive and negative instructor performance or distinguish in such low importance that they can be ignored. These variables are Q1 ''how frequently did you attend this course,'' Q2 ''how frequently did the instructor attend this course,'' Q3 ''did the instructor come to class on time,'' Q9 ''there was a good balance between theory and application,'' Q10 ''the instructor was effective in presenting the material in lectures and discussions,'' Q11 ''subjects were not well organized and did not follow each other,'' Q19 ''this was an easy AA course,'' and Q20 ''the assignments/exams increased the ability to think creatively.''

V. CONCLUSION

Data mining techniques are applied in higher education more and more to give insights to educational and administrative problems in order to increase the managerial effectiveness. However, most of the educational mining research focuses on modeling student's performance [26], [29], [41], [42]. In this paper, data mining is utilized to analyze course evaluation questionnaires. Here, the most important variables that separate ''satisfactory'' and ''not satisfactory'' instructor performances based on students' perception are found. Hopefully, these can help instructors to improve their performances.

In addition, irrelevant variables that do not differentiate ''satisfactory'' and ''not satisfactory'' instructor performances are also listed. Different dimensions of course and instructor effectiveness are measured with course evaluation questionnaires in higher education institutions and these findings may be used to improve measurement instruments.

Furthermore, data mining accurately classifies ''satisfactory'' and ''not satisfactory'' instructor performances. Four different classification techniques –decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis– and seven different classifiers are used and all have performance measures approximately 90% and above in test dataset. This finding indicates the effectiveness of using data mining techniques in course evaluation data and higher education mining. It is hoped this study can contribute to the literature in two major areas: data mining, and higher education.

As a result, the contributions of this study to the literature can be summarized as follows: firstly, effectiveness and expressiveness of data mining techniques, specifically decision tree algorithms, boosting, SVM, ANN, and DA in higher educational mining are presented over a dataset from the daily life. Secondly, using the findings of the variable importance analysis for the classifiers, it is shown that there are many possible improvement areas in the design of the measurement instruments used in instructors' performance evaluation.

REFERENCES

[1] A. M. Abaidullah, N. Ahmed, and E. Ali, ''Identifying hidden patterns in students' feedback through cluster analysis,'' *Int. J. Comput. Theory Eng.*, vol. 7, no. 1, pp. 16–20, 2015.
[2] N. Delavari, S. Phon-Amnuaisuk, and M. R. Beikzadeh, ''Data mining application in higher learning institutions,'' *Inform. Edu.-Int. J.*, vol. 7, no. 1, pp. 31–54, 2007.
[3] M. Goyal and R. Vohra, ''Applications of data mining in higher education,'' *Int. J. Comput. Sci. Issue*, vol. 9, no. 2, pp. 113–120, 2012.
[4] J. Luan, ''Data mining and its applications in higher education,'' in *New Directions for Institutional Research*, vol. 113. New York, NY, USA: Wiley, 2002.
[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Waltham, MA, USA: Morgan Kaufmann, 2012.
[6] J. Luan, ''Data mining and knowledge management in higher education - potential applications,'' in *Proc. AIR Forum*, Toronto, ON, Canada, 2002, pp. 1–16.
[7] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, ''Predicting student performance: An application of data mining methods with an educational Web-based system,'' in *Proc. 33rd Annu. IEEE Frontiers Edu.*, vol. 1. Nov. 2003, p. T2A-13.

[8] V. Kumar and A. Chadha, "An empirical study of the applications of data mining techniques in higher education," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 3, pp. 80–84, 2011.
[9] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, pp. 135–146, Jul. 2007.
[10] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 63–69, 2011.
[11] S. Calkins and M. Micari, "Less-than-perfect judges: Evaluating student evaluations," *NEA Higher Edu. J.*, pp. 7–22, Fall 2010.
[12] J. Sojka, A. K. Gupta, and D. R. Deeter-Schmelz, "Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences," *College Teach.*, vol. 20, no. 2, pp. 44–49, 2002.
[13] L. Coburn. (1984). *Student Evaluation of Teacher Performance, ERIC/TME Update Series*. [Online]. Available: http://ericae.net/edo/ED289887.htm
[14] S. A. Radmacher and D. J. Martin, "Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis," *J. Psychol.*, vol. 135, no. 3, pp. 259–269, 2001.
[15] S. M. Hobson and D. M. Talbot, "Understanding student evaluations: What all faculty should know," *College Teach.*, vol. 49, no. 1, pp. 26–32, 2001.
[16] D. L. Crumbley and E. Fliedner, "Accounting administrators' perceptions of student evaluation of teaching (SET) information," *Quality Assurance Edu.*, vol. 10, no. 4, pp. 213–222, 2002.
[17] S.-H. Liaw and K.-L. Goh, "Evidence and control of biases in student evaluations of teaching," *Int. J. Edu. Manage.*, vol. 17, no. 1, pp. 37–43, 2003.
[18] H. C. Koh and T. M. Tan, "Empirical investigation of the factors affecting SET results," *Int. J. Edu. Manage.*, vol. 11, no. 4, pp. 170–178, 1997.
[19] K. Mckinney, "What do student ratings mean?" *Nat. Teach. Learn. Forum*, vol. 7, no. 1, pp. 1–4, 1997.
[20] W. W. Timpson and D. Andrew, "Rethinking student evaluations and the improvement of teaching: Instruments for change at the University of Queensland," *Stud. Higher Edu.*, vol. 22, no. 1, pp. 55–66, 1997.
[21] J. E. Whitworth, B. A. Price, and C. H. Randall, "Factors that affect college of business student opinion of teaching and learning," *J. Edu. Bus.*, vol. 77, no. 5, pp. 282–289, 2002.
[22] M. Ahmadi, M. M. Helms, and F. Raiszadeh, "Business students' perceptions of faculty evaluations," *Int. J. Edu. Manage.*, vol. 15, no. 1, pp. 12–22, 2001.
[23] C. R. Emery, T. R. Kramer, and R. G. Tian, "Return to academic standards: A critique of student evaluations of teaching effectiveness," *Quality Assurance Edu.*, vol. 11, no. 1, pp. 37–46, 2003.
[24] G. Solinas, M. D. Masia, G. Maida, and E. Muresu, "What really affects student satisfaction? An assessment of quality through a university-wide student survey," *Creative Edu.*, vol. 3, no. 1, pp. 37–40, 2012.
[25] P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching the state of the art," *Rev. Edu. Res.*, vol. 20, no. 10, pp. 1–45, 2013.
[26] A. Peña-Ayala, "Review: Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, 2014.
[27] N. Hajizadeh and M. Ahmadzadeh, "Analysis of factors that affect students' academic performance—Data mining approach," *Int. J. Adv. Stud. Comput. Sci. Eng.*, vol. 3, no. 8, pp. 1–4, 2014.
[28] O. K. Oyedotun, S. N. Tackie, E. O. Olaniyi, and A. Khashman, "Data mining of students' performance: Turkish students as a case study," *Intell. Syst. Appl.*, vol. 7, no. 9, pp. 20–27, 2015.
[29] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Edu. Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
[30] R. Quinlan. (2004). *C5.0: An Informal Tutorial*. [Online]. Available: http://www.rulequest.com/see5-unix.html
[31] L. Breiman, J. H. Friedman, R. A. Olshen, and P. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth International Group, 1984.
[32] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.
[33] L. Breiman, "Arcing classifier (with discussion and a rejoinder by the author)," *Ann. Statist.*, vol. 26, no. 3, pp. 801–849, 1998.
[34] C. Cortes and V. Vapnik, *Machine Learning*. Boston, MA, USA: Kluwer, 1995, pp. 273–297.
[35] S. P. Curram and J. Mingers, "Neural networks, decision tree induction and discriminant analysis: An empirical comparison," *J. Oper. Res. Soc.*, vol. 45, no. 4, pp. 440–450, 1994.
[36] F. J. Hair, Jr., C. W. Black, J. B. Babin, and E. R. Anderson, *Multivariate Data Analysis*, 7th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2010, pp. 335–438.
[37] J. M. Lattin, D. J. Carroll, and P. E. Green, *Analyzing Multivariate Data* (Duxbury Applied Series). Pacific Grove, CA, USA: Thomson Learning Inc., 2003, pp. 426–473.
[38] G. C. J. Fernandez, "Discriminant analysis, a powerful classification technique in data mining," in *Proc. 27th Annu. SAS Users Group Int. Conf. (SUGI)*, Orlando, FL, USA, Apr. 2002, pp. 1–9, paper 247-27.
[39] S. Sharma, *Applied Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 1996, pp. 237–286. [Online]. Available: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471310646.html
[40] R. W. Klecka, *Discriminant Analysis*. Newbury Park, CA, USA: Sage Publications, Inc., 1980.
[41] J. Zimmerman, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance," *J. Edu. Data Mining*, vol. 7, no. 3, pp. 151–176, 2015.
[42] F. D. Kentli and Y. Sahin, "An SVM approach to predict student performance in manufacturing processes course," *Energy, Edu., Sci. Technol. B*, vol. 3, no. 4, pp. 535–544, 2011.

MUSTAFA AGAOGLU received the B.Sc. degree in physics from Bogazici University, Istanbul, Turkey, in 1999, the M.Sc. degree in computer engineering from Marmara University, Istanbul, Turkey, in 2003, and the Ph.D. degree in informatics from Marmara University, Istanbul, in 2009.

He was a Research Assistant with the Faculty of Engineering, Marmara University, from 2000 to 2009, where he has been an Assistant Professor since 2009. His research interest includes data mining, database systems, management information systems, and project management.