DE GRUYTER OPEN

European Journal of Open, Distance and e-Learning
Vol. 17 / No. 1 – 2014
DOI: 10.2478/eurodl-2014-0008

PREDICTING DROPOUT STUDENT: AN APPLICATION OF DATA MINING METHODS IN AN ONLINE EDUCATION PROGRAM

Erman Yukselturk [eyukselturk@gmail.com], Department of Computer Education and Instructional Technology, Kirikkale University, 71450, Kirikkale, Turkey
Serhat Ozekes [serhat.ozekes@uskudar.edu.tr], Department of Computer Engineering, Uskudar University, 34662, Uskudar Istanbul, Turkey
Yalın Kılıç Türel [yturel@gmail.com], Department of Computer Education and Instructional Technology, Fırat University, 23199, Elazığ, Turkey

Abstract

This study examined the prediction of dropouts through data mining approaches in an online program. The subject of the study was selected from a total of 189 students who registered to the online Information Technologies Certificate Program in 2007-2009. The data was collected through online questionnaires (Demographic Survey, Online Technologies Self-Efficacy Scale, Readiness for Online Learning Questionnaire, Locus of Control Scale, and Prior Knowledge Questionnaire). The collected data included 10 variables, which were gender, age, educational level, previous online experience, occupation, self efficacy, readiness, prior knowledge, locus of control, and the dropout status as the class label (dropout/not). In order to classify dropout students, four data mining approaches were applied based on k-Nearest Neighbour (k-NN), Decision Tree (DT), Naive Bayes (NB) and Neural Network (NN). These methods were trained and tested using 10-fold cross validation. The detection sensitivities of 3-NN, DT, NN and NB classifiers were 87%, 79.7%, 76.8% and 73.9% respectively. Also, using Genetic Algorithm (GA) based feature selection method, online technologies self-efficacy, online learning readiness, and previous online experience were found as the most important factors in predicting the dropouts.

Keywords: Educational data mining, student dropout prediction, k-nearest neighbour, decision tree, Naive Bayes, neural network.

Introduction

The recent Internet and Web technologies help higher educational institutions to design and offer online educational opportunities to meet the student and adult needs, such as, convenience and flexibility (Yukselturk, 2009). With the help of these technologies, the number of online degree programs and courses has significantly increased in the new century (Allen & Seaman, 2007). Despite the increasing in the number of online courses and programs, online learning suffers from several problems. One of the main concerns of online learning is high dropout rate. The dropout rates for online education are generally higher than conventional education. Many students are easily leaving online learning courses and programs or finishing without satisfaction (Carr, 2000; Inan et al., 2009; Kotsiantis et al., 2003; Lykourentzou et al., 2009; Willging & Johnson, 2004). Student retention and attrition in online education has been actively researched for a long time. Several researchers emphasized that limiting dropout and keeping learners online is essential in online learning since dropout rates are considered as a quality indicator of online

education (Lykourentzou et al., 2009; Willging & Johnson, 2004). Also, if potential dropouts could be identified early, prevention might be possible, and so keep more successful and satisfied students online (Inan et al., 2009; Lykourentzou et al., 2009). In the literature, when analyzing related researches, their topics are mainly associated with the factors affecting student persistence/dropout, and analyzing interventions aiming to improve retention. Retention researches focuses on investigation of student attrition behaviours, analyses of graduation rate, examination of persistence patterns, and explanations of the psychosocial dynamics related to retention (Inan et al., 2009; Lykourentzou et al., 2009; Willging & Johnson, 2004; Yukselturk & Inan, 2006). Also, models and instruments to assess, predict, and enhance student retention have been also developed in several studies (Berge & Huang, 2004; Simpson, 2004).

On the other hand, the ability to predict dropouts and improve retention is a still complex issue that involves the number of inter-correlated and distinct factors. These factors might vary in different context and student retention is going to get even worse in several open and distance learning. Also, some inventions and models discussed in several studies have varied weaknesses. For example, that could be small scale and context-dependent (Gibbs, 2003). In addition, due to the increase in the number of online students, the vast quantities of data have been gathered related to personal information about the students' academic achievements, interactions, and their dropouts day by day in higher institutions. Therefore, researchers, instructors and administrators seek effective methods to extract meaningful knowledge from the large data sets. One of the major analytical methods can be used for this purpose is data mining (Beikzadeh et al., 2008; Benoît, 2002; Minaei-Bidgoli et al., 2004; Romero, Ventura & García, 2008). It is simply defined as the nontrivial process of identifying valid, potentially useful, novel, and eventually understandable patterns in data (Fayyad et al., 1996). The use of data mining in education has grown in recent years for several reasons: a considerable increase in the amount of data, technological advances in computer sciences, and well-developed of tools for analyses (Barker & Siemens, in press).

Data mining includes several types of tasks. One of them is the classification that is the task of delineating known entities to apply to new data. It is a supervised learning algorithm in which entities are arranged into classes based on common characteristics of the training set which is previously labelled (Minaei-Bidgoli et al., 2004; Romero, Ventura & García, 2008). The most popular classification algorithms, also used in this study, are decision trees, neural networks, k-nearest neighbours, Naive Bayes based on applying Bayes' theorem and genetic algorithm. In this study, these algorithms are presented in detail in the discussion of data analysis under the method section.

Background

Data mining (DM), which provides valid and interpretable results to researchers, is becoming an essential way to transform data into information in a wide range of areas, such as marketing, banking, surveillance, and fraud detection (Benoît, 2002). In education, DM has been applied to data retrieved from the implementation of different instructional modalities including, particularly, computer-based and web-based education as well as traditional (face-to-face) education. Recently, to extract unexpected knowledge and discover various patterns in large data sets from especially e-learning systems for the purpose of solving educational problems, "educational data mining" has been used as a common term (Lile, 2011). Since computer and web-based educational systems can record vast amounts of student profile data onto log files and databases, DM techniques can be applied to find interesting and unanticipated relationships among attributes of students, teaching and learning strategies, and assessments (Beikzadeh et al., 2008; Romero, Ventura & García, 2008).

ISSN 1027-5207

Although the common goal of traditional statistical methods and DM techniques is to retrieve the most valuable information from the collected data, DM techniques promise a great potential in terms of knowledge discovery since they embrace various disciplines such as statistics, artificial intelligence, and machine learning as new and popular paradigms (Zhao & Luan, 2006). As well as its potential to deal with a huge amount of data set, DM allows us to conduct a predictive analysis that cannot easily be associated with a particular theory used for identifying variables regarding the context of the study (Zhao & Luan, 2006). McCarthy and Earp (2009) suggest that a researcher can conduct a logistic regression analysis to recognize and describe characteristics regarding reporting errors but that analysis entails hypotheses concerning the characteristics of the errors and predefined interaction effects. However, as a DM technique, classification trees (also called as decision trees) can automatically recognize substantial interaction effects without any predetermined hypothesis in regard to target variable relationships (McCarthy & Earp, 2009).

As an example from current studies in the literature, Hämäläinen et al. (2004) constructed a Bayesian network to describe the student's learning process. The aim of this study was to classify students to give them differentiated guiding in accordance with their skills and other characteristics (Hämäläinen et al., 2004). In another study about intelligent tutoring systems, Beck and Woolf (2000) constructed a learning agent for high-level student modelling with machine learning. The agent learned predicting the probability the student's next response would be correct, and how long it would take the student to make that response. They used linear regression to predict observable variables. Similarly, Chen et al. (2000) applied decision tree algorithm and data cube technology from web log portfolios for managing classroom processes, and Talavera and Gaudioso (2004) proposed mining student data using clustering to discover patterns reflecting user behaviours. Furthermore, Su et al. (2006) proposed a learning portfolio mining (LPM) approach including four phases: user model definition phase, learning pattern extraction phase, decision tree construction phase, activity tree generation phase. According to the results, generated personalized activity trees with sequencing rules are workable and beneficial for learners. Minaei-Bidgoli et al. (2003) analyzed the grades and several features of 227 online students via several DM techniques and researchers reported that they got the optimum results by Decision Tree and Neural Network with high accuracy rates: 94% for binary, 72% for 3-classes, and 62% for 9-classes. In a similar research study, Kotsiantis et al. (2003) examined 510 online students' demographics such as sex, age, and occupation as well as their assignments by various DM algorithms and obtained the best results from Naive Bayes method with a 74% accuracy rate. Cortez and Silva (2008) used a 37-item questionnaire to collect data from 788 public school students in Portugal, which includes demographics, social/emotional, and school-related variables regarding student performance. In that study, four DM methods including Decision Trees, Random Forest, Neural Networks, and Support Vector Machines were tested to predict students' Mathematics and Portuguese grades and good accuracy rates were obtained by performed DM methods. For the purpose of investigation of web-based teaching and learning, Zang and Lin (2003) used a questionnaire consisted of 37 items and five sections including students' demographics and overall perspectives about an online course. In that study, researchers used boosting Algorithms as a DM technique to predict students' academic success and found that students who post questions and answers, and surf Internet tend to get high scores from the online courses. Durfee, Schneberger, and Amoroso (2007) analyzed the results of a 32-item computer-based learning (CBA) survey completed by over 550 undergraduate students. After applying Principle Component Analysis to see main dimensions of the data, researchers used the self-organizing map as a popular neural network DM technique to determine the related characteristic of data as clusters. In another study, Romero and Ventura (2007) analyzed several applications of data mining to traditional educational systems, web-based courses, well-known learning content management systems, and adaptive and intelligent web-based educational systems from 1995 to 2005. They concluded that data mining techniques can be applied in several

ISSN 1027-5207

areas, such as, statistics and visualization, clustering, classification and outlier detection, association rule mining and pattern mining, and text mining. McCarthy and Earp (2009) state that there are few examples of using DM techniques with the data collected via surveys and questionnaires. For instance, Scime and Murray (2007) worked on the exit poll data by means of classification trees method to build frameworks predicting likely voters. Schouten and de Nooij (2005) also used the same DM technique to determine post-stratification weights for participants in a survey regarding household living conditions in Netherlands.

Recently, researchers also have utilized several data mining techniques to predict student dropout or success in higher education level. For instance, Dekker et al. (2009) analyzed the results of the educational data mining case study aimed at predicting the Electrical Engineering students drop out. Their experimental results showed that rather simple and intuitive classifiers (i.e. decision trees) gave a useful and accurate result. Another study was conducted to deal with student dropout problem in Hellenic Open University by Kotsiantis, Pierrakeas, and Pintelas (2003). They suggested that the Naive Bayes algorithm is the most appropriate technique to use for prediction of student dropout based on the comparison of six algorithms. Another study conducted by Superby, Vandamme, and Meskens (2006) aims to determine factors effecting the success and drop-outs of university students using various DM techniques such as neural networks, discriminant analysis, decision trees, and random forests. To overcome individual technique inaccuracies in identifying student dropouts, Lykourentzou et al. (2009) used three different machine learning techniques, namely, neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP (fuzzy logic and adaptive response theory map) for dropout prediction in e-learning courses. They maintained that the most successful technique in predicting students' dropout is the decision scheme. Similarly, Herzog (2006) compared a variety of machine learning techniques to predict student dropouts. In this study, he analyzed the predictive accuracy of neural networks, decision trees and multinomial logistic regression, over the problem of predicting college freshmen retention. The results indicated that all of the examined methods achieved similar correctness ratios (Herzog, 2006).

To summarize, the field of data mining is concerned with finding new patterns in large amounts of data. Widely used in business, data mining also has several application areas in education. One of the most useful DM tasks in online learning is the ability to classify a student's retention and attrition. There are different educational objectives for using classification, such as: exploring students' reactions towards a certain instructional strategy, discovering students' characteristics and so forth (Romero et al., 2008). The aim of the study is to examine dropout prediction with the help of student personal characteristics using data mining techniques in an online program. The main focus of this study is not only to examine the effects of student personal characteristics on dropout, but also, to attempt performing various data mining techniques for such analyses. Four different classifiers, which were based on k-Nearest Neighbour (k-NN), Decision Trees (DT), Naive Bayes (NB) and Neural Networks (NN), were trained and tested using 10-fold cross validation. The prediction performances of k-NN, DT, NB and NN classifiers were compared by showing the prediction results and plotting ROC (Receiver Operating Characteristica graph of sensitivity, y-axis, versus specificity, x-axis) curves. Also, Genetic Algorithm (GA) was applied to find an optimal set of feature weights that are the most important factors in dropout prediction. As such, we provide a good sample and an illustration of using these various techniques for the analysis of students' dropouts based on the surveyed data, which has potential to contribute to existing literature in this context.

ISSN 1027-5207

Method

In this section, participants, data collection and analysis of this study were described in a detailed way under three subsections.

Subject of the study

The data was collected from the students who registered to online Information Technologies Certificate Program (ITCP) offered/delivered by a government university in the capital city of Turkey. The main aim of this program is to train participants in information technology to meet demands in the field of computer and information technologies. The program, which is still active, consists of eight fundamental courses of the Computer Engineering Department and is comprised of four semesters lasting a total of nine months (Isler, 1998; Yukselturk, 2009).

This online program accepts only students who are studying or graduated from two-year colleges or four-year universities. Also, the students are expected to be computer literate and have an intermediate level of English. A total of 189 students who registered to this program in 2007-2009 were included in this study. The percentages of students who registered and completed the program (N=120) were 63.49% and dropped the program (N=69) were 36.51%.

Table 1 presents the percentages of all participants' demographic characteristics. The number of male students (70%) was greater than the number of female (30%) students, and the students' ages ranged from 20 to 55 with an average of 28. The majority of the online program students were undergraduate and graduate student (MS or PhD students) (60.3%). Nearly half of the students (49.7%) have full-time or part-time jobs and only a few of them (10.5%) have previously been in an online course.

Table 1: The Demographic Characteristics of Participants

Gender	**# of registered participants**	**# of dropout participants**	**percentage of registered participants**	**percentage of dropout participants**
Female	31	26	25.83	37.68
Male	89	43	74.17	62.32
Age				
20-29	99	45	82.50	65.22
30-39	15	19	12.50	27.54
40+	6	5	5.00	7.25
Educational Level				
Graduate	45	30	37.50	43.48
Undergraduate student	59	32	49.17	46.38
Graduate student (MS or PhD student)	16	7	13.33	10.14
Previous Online Experience				
Yes	14	6	11.67	8.70
No	106	63	88.33	91.30
Occupation				
Working	59	35	49.17	50.72
Not working	61	34	50.83	49.28

ISSN 1027-5207

Table 2 shows descriptive statistics of the participants' initial perceptions about online technologies self-efficacy, online learning readiness, locus of control, and prior knowledge at the beginning of the online program. Both registered and dropout participants had a positive perception about online technologies based on their self-efficacy scores. Most participants thought that they were ready for online learning, and also participants` perceptions in terms of their locus of control were generally low. It means that participants had better control of their behaviour and perception for their own life and actions. Moreover, most participants stated that they did not have much knowledge about program courses at the beginning of the online certificate program. Finally, according to the results, the registered participants mean scores were slightly higher than dropout ones (see Table 2).

Table 2: Descriptive Statistics of the Participants' Initial Perceptions about Predictors

	registered participants		**dropout participants**	
	Mean	**Std.**	**Mean**	**Std.**
online technologies self-efficacy (out of 116)	105.5	12.2	103.9	15.5
online learning readiness (out of 52)	43.26	5.31	43.7	4.5
locus of control (out of 23)	8.39	3.9	7.4	4.0
prior knowledge (out of 32)	13.13	4.39	12.3	4.3

Data collection

In this study, five online questionnaires were administered in order to determine participants' characteristics and initial perceptions: Demographic Survey (DS), Online Technologies Self-Efficacy Scale (OTSE), Readiness for Online Learning Questionnaire (ROLQ), Locus of Control Scale (LCS), and Prior Knowledge Questionnaire (PKQ). These selected questionnaires were translated into Turkish and used in previous studies. At the beginning of the online program, a five-hour face-to-face orientation was organized to explain the program and courses to the participants, help them meet each other and with the instructors, explain how to use the web pages, and also mentioned about the questionnaires related to this study. After the orientation, all online questionnaires were administered during the first week of the program. Participants submitted their responses to these online questionnaires and their data was saved on the database server.

DS was used to gather students' demographic information (i.e. age, gender, education level). OTSE was used to measure students' self-efficacy beliefs specific to the online environment. It was originally developed by Miltiadou and Yu (2000) and it is a 29-item Likert scale with five subscales. This scale was translated into Turkish with a high Cronbach alpha level of 0.97 by one of the researchers of this study (Yukselturk, 2009).

ROLQ was used to assess students' readiness for online learning. It was originally developed by McVay (2000) and composed of 13 items, rated by respondents on a Likert scale. This scale was also translated into Turkish with a Cronbach alpha level of 0.76 by Yukselturk (2009).

ISSN 1027-5207

The Internal-External LCS was used to measure students' locus of control orientation. It was originally developed by Rotter (1966). It is a 29 item forced-choice self-report scale with scoring range 0 (internality) to 29 (externality) excluding 6 buffer items. The scale was translated into Turkish and standardized on a Turkish sample (n=532) by Dag (1991). He calculated the Cronbach alpha coefficient of LCS in his study as 0.71.

PKQ was used to assess students' prior knowledge about online program courses. It consisted of eight items and each item was related to each aim of courses which are given in the program. By means of PKQ items, participants were asked to indicate their level of knowledge related to the program courses (Yukselturk, 2009).

The survey dataset comprised of five demographic variables (gender, age, educational level, previous online experience, and occupation), self-efficacy (measure students' self-efficacy beliefs specific to the online environment), readiness (assess students' readiness for online learning), prior knowledge (assess students' prior knowledge about online program courses), locus of control (measure students' locus of control orientation). In addition, the dropout status indicated whether students continue to attend the program or not in this study.

Data analysis

In order to classify the dropout students, four data mining approaches were applied by using MATLAB: k-Nearest Neighbour (k-NN), Decision Tree (DT), Naive Bayes (NB), and Neural Network (NN), respectively. Since these methods are supervised and most-preferred ones in the literature, we have selected them as our main DM methods in this study. The inputs of these classifiers were nine variables of the students which were gender, age, educational level, previous online experience, occupation, self efficacy, readiness, prior knowledge, and locus of control. While selecting these variables, we have prioritized several issues since DM approaches are sensitive. Therefore, the data were organized carefully, then, inaccurate records and extreme outliers were cleaned and removed while data analysis process. On selection of which classification algorithm to use in this study, two steps were carried out respectively. First, we tested the algorithm and then, designated it as the classifier for routine use if its prediction accuracy was satisfactory. The technique, used to calculate a classifier's accuracy, was cross-validation. In k-fold cross-validation, the training set is divided approximately equal-sized k subsets. The classifier is trained k times using the subsets by leaving one out for each time. Therefore, the average error rate of the each subset can be used as estimation for the error rate of the classifier (Kotsiantis, 2007). Thus, the average error rate provides a better solution in terms of the higher reliability of the accuracy rates. In this study, the 10-fold cross-validation results of the classifiers were evaluated since we aimed to increase the sensitivity of the results by the selection of 10 as a k-fold value. Then, genetic algorithm was applied to find an optimal set of feature weights that were the most important factors in predicting dropouts in the online program.

k-Nearest Neighbour classifier (k-NN)

k-NN algorithm is one of the well-known classification methods. It is based on learning by comparing a given test tuple with training tuples that are similar to it. When a new instance is introduced, k-NN finds the k-nearest neighbours of this new instance and determines the label of the new instance by using these *k* instances (Hand, Mannila & Smyth, 2002).

In this study, closeness is defined in terms of a distance metric *d* called Euclidean distance (Han & Kamber, 2006). Although our data set is mostly consisted of categorical variables, each category has a numerical counterpart; thus, we have used the Euclidean distance. To assign a particular class to the test sample, the most common class among the *k* nearest neighbours is

ISSN 1027-5207

used and the unclassified test sample is classified by a majority vote of its neighbours. A good value for *k*, the number of neighbours, was determined experimentally. Starting with *k*=1, we used the 10-fold cross validation technique to estimate the error rate of the classifier. This process was repeated for *k*=10 times and in each iteration by incrementing *k* to allow for one more neighbour. The value *k* was selected as '3' that gave the minimum error rate. According to Hämäläinen and Vinni (2010), this method has several advantages: The accuracy rate of classification can be very satisfying in some cases, there are just two parameters to learn, and the classification is very robust to missing values. However, the selection of distance function *d* might be difficult particularly for educational data sets (Hämäläinen & Vinni, 2010).

Decision Tree classifier (DT)

DT is a powerful and popular classification and prediction technique (Chaudhuri, 1998). Hämäläinen and Vinni (2010) stress that DT is the most common DM technique in the literature. There are several popular decision tree algorithms such as ID3, C4.5, and CART (classification and regression trees). DT is in the form of a tree structure, where each node is either a leaf node (indicating the value of the target class of examples) or a decision node (specifying a test to be carried out on a single attribute value, with one branch and sub-tree for each possible outcome of the test) (Berson, Smith & Thearling, 2000). DTs have many advantages such as very fast classification of unknown records, easy interpretation of small-sized trees, robust structure to the outliers' effects, and a clear indication of most important fields for prediction but DTs are very sensitive to over-fitting particularly in small data-sets (Hämäläinen & Vinni, 2010).

In this study, to generate a decision tree, the C4.5 (Quinlan, 1993) algorithm was used, which is an extension of Quinlan's earlier ID3 algorithm. To construct the tree, entropy measure was used in the determination of nodes. Since the attributes with higher the entropy cause more uncertainty in outcome, they were selected in order of increasing entropy.

Naive Bayes classifier (NB)

A simple probabilistic classifier called as Naive Bayes classifier was also used in student dropout classification. Naive Bayes algorithm as the simplest form of Bayesian network (Domingos & Pazzani, 1997) is one of the easiest algorithms to perform and has very satisfactory accuracy and sensitivity rates (Kotsiantis, Pierrakeas & Pintelas, 2003). The posterior probability of each class, C_i, is obtained by the Naive Bayes classifier using Bayes rule. The classifier makes the simplifying assumption that the attributes, A, are independent given the class, so the likelihood can be obtained by the product of the individual conditional probabilities of each attribute given the class (Flach & Lachiche, 2004). Thus, the posterior probability, $P(C_i \mid A_1,...,A_n)$, can be given by the following equation/assumption:

$$P(C_i \mid A_1,...,A_n) = P(C_i)P(A_1 \mid C_i)...P(A_n \mid C_i)/P(A)$$

This assumption is usually called the *Naive Bayes assumption*, and a Bayesian classifier using this assumption is called the *Naive Bayesian classifier*, often abbreviated to 'Naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class (Flach & Lachiche, 2004).

Neural Network classifier (NN)

The prediction of the student dropouts was also performed by feed-forward NN. It is another inductive learning method grounded on computational models of neurons and their networks as in humans' central nervous system (Mitchell, 1997). NN is a set of connected input/output units where each connection has a distinct weight associated with each other (Kotsiantis, Pierrakeas & Pintelas, 2003). During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class of the input samples (Han & Kamber, 2006).

In this study, the back propagation algorithm was performed for learning on a multilayer feed-forward neural network. The input layer of the network consisted of nine variables of the students. The hidden layer included 50 neurons and the output layer had one neuron, which was determined by our experimental studies.

Genetic Algorithm (GA)

In this study, we have also used the GA as a feature selection method after the classification algorithms. GA is a paradigm developed based on natural selection and Darwin's theory about evolution, and used for solving optimization problems (Romero et al., 2002). Algorithm starts with a random initial set of solutions for the feature subset selection problem. This initial set is called as chromosome. The chromosome was represented by a binary vector of dimension nine, the total number of features. After the initial population is selected randomly, roulette wheel selection, mutation and uniform crossover operations continue until a new generation is replaced with the original generation. It is expected that the new generation would give a better solution than the old one. Solutions are selected according to their fitness to form new solutions (Davis, 1991).

A lower fitness value refers a more optimized solution to the problem, which is minimizing the classification error after 10-fold cross-validation in this study. These steps were repeated until the number of populations, which is 200, is satisfied. Due to the high number of the solution space, we have used the number of population throughout the analyses. Over successive generations, the population evolved toward an optimal solution.

Results

In this study, the classification of dropout students was performed using the data mining models based on k-NN, DT, NB and NN, respectively. 10-fold cross validation was performed to validate each classifier. When 10 tests were completed, the average performance on the tests was used to determine the accuracy of the model developed.

The 10-fold cross validation of the k-NN classifier was performed 10 times in order to find a good k value that was incremented to allow one more neighbour in each iteration. Figure 1 shows the accuracy of k-NN classifier for 10 different k values. As seen in Figure 1 the best accuracy was achieved with 3-NN classifier.

ISSN 1027-5207

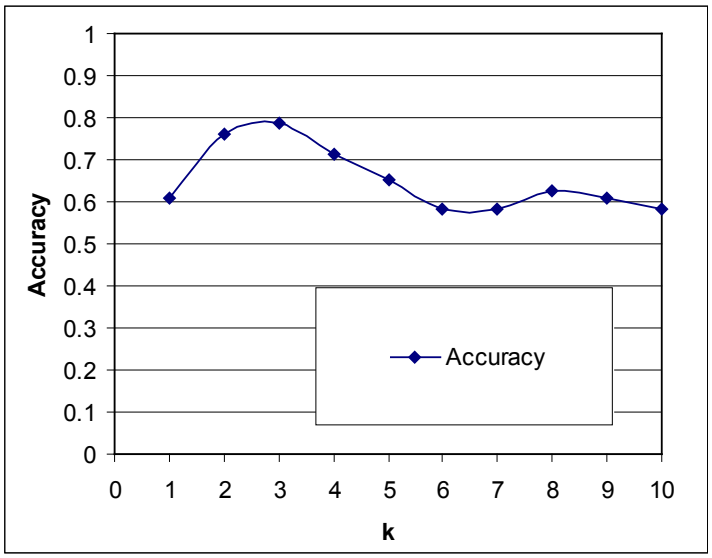

Figure 1. 10-fold cross validation accuracy levels of k-NN classifier for different k values

The detection sensitivities of 3-NN, DT, NN, and NB classifiers were 87%, 79.7%, 76.8%, and 73.9% respectively. Receiver Operating Characteristic (ROC) curve is also used for the evaluation of classification algorithms.

In ROC analysis, true positive and false positive rates describe the performance of the model independently of the class distribution (Flach, 2003). ROC is a plot of the sensitivity (true positive rate-TP) as the function of false positive rate (FP) (1-specificity). A test with perfect discrimination has a ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Area under ROC curve (AUC) is the measure of separability of classification functions and calculated based on the following formula (Vuk & Curk, 2006):

$$A_{ROC} = \int_0^1 \frac{TP}{P} \; d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP \; dFP$$

In Figure 2, the ROC curves showing the 10-fold cross validation performances of classification methods are seen. The AUCs of 3-NN, DT, NB, and NN were 0.866, 0.861, 0.574, and 0.526 respectively.

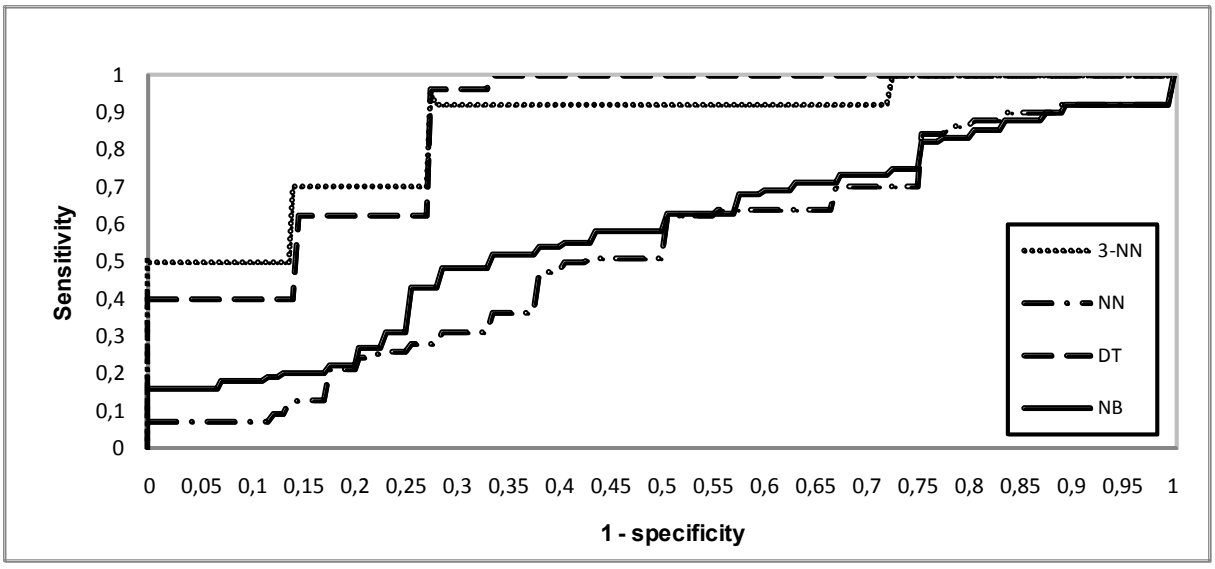

Figure 2. ROC curves for 3-NN, NN, DT, and NB classifiers

ISSN 1027-5207

According to the 10-fold cross-validation performances of the classifiers, sensitivity was significantly the best for 3-NN classifier compared to NN, DT, and NB. There was not a big difference in sensitivities between NN, DT, and NB classifiers. Also, AUC was better for 3-NN and DT classifiers than NN and NB classifiers. The reasons for the failure of these NB and NN classifiers can be given as follows:

- Bayes classifier considers the class information attributes as independent while those attributes are dependent in real life.
- In Bayes classifier, the relationship between the variables cannot be modelled.
- Although NN's parameters (i.e., learning rates and no. of hidden layer neurons) were randomly selected, they were not satisfactory.

Finally, GA based feature selection method was used to determine the significant factors in predicting the online program dropouts. Each classifier was used with GA to calculate the fitness values using their classification errors after 10-fold cross validation. The features selected for each classifier after the GA are listed in Table 3. According to the results, three variables, which are online technologies self-efficacy, online learning readiness, and previous online experience, were found as the most important factors in predicting the dropouts.

Table 3: The Most Important Factors based on Data Mining Methods in Predicting Dropouts

Data Mining Techniques	**3-NN**	**DT**	**NN**	**NB**
Variables	Readiness	Self efficacy	Self efficacy	Self efficacy
	Prev. online experience	Readiness	Readiness	Gender
	Gender	Prev. online experience	Prev. online experience	Age
			Locus of control	Prior knowledge

According to whether the target criteria (threshold) determined by each one of the four DM methods above are met, results in Table 3 were found. For example, for the 3-NN method, results were deducted based on the minimum Euclidean distance. For DT, we took the entropy measure into account as a threshold. In similar, we determined 10^{-4} as a goal performance for NN. At the end of the training process, the value of mean square error (MSE) is under the determined goal-performance. Finally, we took the highest value of the posterior probability of each class as threshold value for NB. Those criteria provide the reliability of obtained results.

Conclusion and future work

Predicting dropout student is an important and challenging task for universities, policymakers, and educators especially in online learning. Therefore, this study examined whether the use of data mining techniques can be helpful in dealing with this problem in an online program. In order to classify the dropout student, the most common data mining approaches were applied based on k-Nearest Neighbour (k-NN), Decision Tree (DT), Naive Bayes (NB) and Neural Networks (NN). These four different classification algorithms were trained and tested using 10-fold cross-validation technique. Although there was not a big difference in sensitivities among these four classifier algorithms, the 3-NN (87%) and DT (79.7%) classifiers were more sensitive. Also, these sensitive scores were promising results for predicting dropout student in the online program dataset. This finding was in line with previous research results related to classifier algorithms predicting dropouts. For example, Kotsiantis, Pierrakeas, and Pintelas (2003) mentioned that the learning algorithms predict dropout of new students with satisfying accuracy

ISSN 1027-5207

and; thus, become a useful tool in an attempt to prevent and reduce dropouts. In their study, the accuracy reaches 63% in the initial predictions based only on demographic data of the students and exceeds 83% before the middle of the academic period. In another study, Dekker et al. (2009) found that DT gave a useful result with accuracies between 75% and 80% in predicting student dropout. As a result, data mining techniques could be a useful tool with satisfying accuracy to be able to predict different paths of students who leave the programs before completing all requirements.

Researchers stated that various student demographics factors have been used to predict occurrences of dropouts in the literature (Inan, Yukselturk & Grant, 2009; Lykourentzou et al., 2009; Willging & Johnson, 2004). Similarly, according to the results of GA based feature selection method in this study, participants' initial perceptions about self-efficacy beliefs specific to the online environment, readiness for online learning and prior knowledge about course contents were important predictive variables that related to student dropouts. It was proved that students might have some necessary characteristics and skills before online learning, or, the instructors might provide these students with special attention to reduce or prevent dropouts during online learning, as exposed by previous studies (Inan, Yukselturk & Grant, 2009; Lykourentzou et al., 2009). With the help of data mining techniques, this type of results could also help people who deal with online courses and programs orient their students to the kind of skills and abilities they will need to function well in online programs.

It is apparent that using data mining techniques in educational studies promises opportunities for educators and researchers to achieve more useful knowledge and more interesting relationships between variables from the large data sets. In other words, through data mining, researchers could discover which behaviours and decisions lead to learner success, identify learners who are at risk of dropping out or failing, personalize and adapt learning content and instruction to meet individual needs, and improve and optimize the use of institutional support structures to assist learners (Baker, 2010). Moreover, considering student apathy and incomplete surveys in regard to current data collection methodologies such as paper-based surveys, it can be more effective and helpful for educational researchers to collect comprehensive data from online learning environments and then, to apply data mining techniques as suggested by Black, Dawson and Priem (2008). It seems that data mining techniques offer a number of advantages for mainly the studies in education and also in other social sciences. However, there is a paucity of existing literature in regard to the use of data mining techniques despite rapidly growing and easily achievable data in online settings and the urgent need for alternative and viable analyses approaches for educational studies. To fill a void that currently exists in the research, the results of this type of study might be potential example for researchers. Also, researchers, who have limited technical background, may have difficulty in applying sophisticated data mining tools and algorithms to their own data without an inter-disciplinary collaboration. As Hung and Zhang (2008) stressed, it is also essential to develop more user-friendly and functional data mining tools for educators.

To summarize, it seems that more students will have entered online learning courses and programs in the following years; therefore, more data would be gathered in databases about several kinds of student information. We need more research to extract and generalize meaningful knowledge from these student information using data mining efforts in an online learning domain. The main restrictions of this study were based only single experiment and on a limited sample of students. Other researchers could replicate this type of study with larger sets of student data from different online degrees and certificate programs. In addition, further studies might include changing and adding the variables, applying other algorithms, and modifying the pre-processing methods. Also, triangulation of the research method with qualitative data (i.e. interviewing with dropouts) may help researchers validate and interpret the results of each data

ISSN 1027-5207

mining technique by seeing the multidimensional picture of the problem. Besides, these types of studies could be applied to other disciplines in order to achieve the highest possible prediction accuracy.

References

1. Allen, I.E. and Seaman, J. (2007). *Online nation: Five years of growth in online learning.* Needham, MA: Sloan Consortium.
2. Baker, R.S.J.D. (2010). Data Mining for Education. In B. McGaw, P. Peterson, E. Baker (eds.), *International Encyclopaedia of Education (3rd edition)*, (pp. 112-118). Oxford, UK: Elsevier
3. Baker, R. and Siemens, G. (in press). Educational data mining and learning analytics. To appear in Sawyer, K. (ed.), *Cambridge Handbook of the Learning Sciences: 2nd Edition.*
4. Beck, J. and Woolf, B.P. (2000). High-level student modeling with machine learning. In G. Gauthier, C. Frasson & K. VanLehn (eds.), *Proceedings of Fifth International Conference on Intelligent Tutoring Systems,* (pp. 584–593). Berlin: Springer-Verlag Berlin & Heidelberg GmbH & Co. K.
5. Beikzadeh, M.R.; Phon-Amnuaisuk, S. and Delavari, N. (2008). Data mining application in higher learning institutions. In *International Journal of Informatics in Education, 7(1),* (pp. 31-54).
6. Benoît, G. (2002). Data mining. In *Annual Review of Information Science and Technology, 36,* (pp. 265–310).
7. Berge, Z. and Huang, Y. (2004). A Model for Sustainable Student Retention: A Holistic Perspective on the Student Dropout Problem with Special Attention to e-Learning. In *DEOSNEWS, 13(5),* Retrieved July 29,2011, http://www.ed.psu.edu/acsde/deos/deosnews/deosnews13_5.pdf
8. Berson, A.; Smith, S. and Thearling, K. (2000). *Building Data Mining Applications for CRM.* New York: McGraw-Hill Professional Publishing.
9. Black, E.W.; Dawson, K. and Priem, J. (2008). Data for free: using LMS activity logs to measure community in online courses. In *The Internet and Higher Education, 11(2),* (pp. 65-70).
10. Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. In *The Chronicle of Higher Education, 46(23),* (pp. A39-A41).
11. Chaudhuri, S. (1998). Data Mining and Database Systems: Where is the Intersection? In *IEEE Bulletin of the Technical Committee on Data Engineering, 21(1)*, (pp. 4-8).
12. Chen, G.; Liu, C.; Ou, K. and Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. In *Journal of Educational Computing Research, 23(3),* (pp. 305–332).
13. Cortez, P. and Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito & J. Teixeira (eds.), *EUROSIS*, (pp.5-12).
14. Davis, L. (1991). *Handbook of Genetic Algorithms.* New York, NY: Van Nostrand Reinhold
15. Dag, I. (1991). The reliability and validity study of Rotter's IE/LOC scale for university students. In *Turkish Journal of Psychiatry, 7(26),* (pp. 10-16).
16. Dekker, G.W.; Pechenizkiy, M. and Vleeshouwers, J.M. (2009). Predicting student drop out: A case study. In T. Barnes, M. Desmarais, C. Romero & S. Ventura (eds.), *Proceedings of the 2nd International Conference on Educational Data Mining,* EDM 2009, Retrieved July 29, 2011, from http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf
17. Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. In *Machine Learning, 29,* (pp. 103-130).

ISSN 1027-5207

18. Durfee, A.; Schneberger, S. and Amoroso, D.L. (2007). Evaluating students' computer-based learning using a visual data mining approach. In *Journal of Informatics Education Research, 9(1),* (pp. 1-28).

19. Fayyad, U.M.; Pitatesky-Shapiro, G.; Smyth, P. and Uthurasamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, Cambridge.

20. Flach, P. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In T. Fawcett & N. Mishra (eds.), *Proceedings 20th International Conference on Machine Learning (ICML'03),* (pp. 194-201). AAAI Press.

21. Flach, P. and Lachiche, N. (2004). Naive Bayesian classification of structured data. In *Machine Learning, 57(3),* (pp. 233-269).

22. Gibbs, M.R. (2003). Knowledge Sharing and Socialization in Distributed Communities of Practice. In R.M. Verburg & J.A. De Ridder (eds.), *Knowledge Sharing Under Distributed Circumstances,* Amsterdam: Netherlands Organization for Scientific Research.

23. Hämäläinen, W. and Vinni, M. (2010). Classifiers for educational technology. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (eds.), *Handbook of Educational Data Mining,* (pp. 54-74). CRC Press.

24. Hämäläinen, W.; Suhonen, J.; Sutinen, E. and Toivonen, H. (2004) Data mining in personalizing distance education courses. In *World conference on open learning and distance education.* Retrieved July 29, 2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.5378&rep=rep1&type=pdf

25. Han, J. and Kamber, M. (2006). *Data Mining Concepts and Techniques.* The Morgan Kaufmann Series in Data Management Systems, 2nd Edition. San Francisco: Elsevier Inc.

26. Hand, D.; Mannila, H. and Smyth, P. (2002). *Principles of data mining.* Cambridge, Massachussetts, USA: MIT Press.

27. Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. In *New Directions for Institutional Research, 2006(131),* (pp. 17–33).

28. Hung, J. and Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. In *MERLOT Journal of Online Learning and Teaching, 4(4),* (pp. 426-437).

29. Inan, F.A., Yukselturk, E. and Grant, M.M. (2009). Profiling potential dropout students by individual characteristics in an online certificate program. In *International Journal of Instructional Media, 36(2),* (pp. 163-176).

30. Isler, V. (1998). *Distance Education Experiences of the Middle East Technical University.* Paper presented at MEDISAT-EUREKA: Joint Workshop: Internet as a Medium for Innovation and Technology Development in Eastern Mediterranean, Tubitak-Bilten & EU/INCO-DC, Ankara, Turkey.

31. Kotsiantis, S.B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In *Informatica, 31(3),* (pp. 249-268).

32. Kotsiantis, S.; Pierrakeas, C. and Pintelas, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems,* (pp. 267–274).

33. Lile A. (2011). Analyzing E-Learning Systems Using Educational Data Mining Techniques. In *Mediterranean Journal of Social Sciences, 2(3),* (pp. 403-419). DOI: 10.5901/mjss.2011.v2n3p403

34. Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G. and Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. In *Computers & Education, 53(3),* (pp. 950-965).

35. McCarthy, J.S. and Earp, M.S. (2009). *Who makes mistakes? Using data mining techniques to analyze reporting errors in total acres operated.* National Agricultural Statistics Service, RDD Research Report Number RDD-09-02. Retrieved, January 21, 2012 from http://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/data-mining-reporting-errors.pdf

36. Mcvay, M. (2000). *Developing a Web-based distance student orientation to enhance student success in an online Bachelor's degree completion program.* Unpublished practicum report presented to the Ed.D. Program, Nova Southeastern University, Florida.

37. Mitchell, T. (1997). *Machine Learning.* New York: McGraw Hill.

38. Miltiadou, M. and Yu, C.H. (2000). *Validation of the online technologies self-efficacy survey (OTSES).* (ERIC Document Reproduction Service No. ED. 445672).

39. Minaei-Bidgoli, B.; Kashy, D.; Kortemeyer, G. and Punch W. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. In *Proceeding of IEEE Frontiers in Education,* (pp. 13–18). Colorado, USA.

40. Minaei-Bidgoli, B.; Kortemeyer, G. and Punch, W.F. (2004). *Enhancing online learning performance: An application of data mining methods.* Paper presented at the 7th IASTED International Conference on Computers and Advanced Technology in Education (CATE 2004), Retrieved July 29, 2011, from http://www.lon-capa.org/papers/Behrouz_CATE2004.pdf

41. Quinlan, J.R. (1993). *C4.5: Programs for machine learning.* San Francisco, CA.: Morgan Kaufmann Publishers.

42. Romero, C. and Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. In *Expert Systems with Applications, 33(1),* (pp. 135-146).

43. Romero, C.; Ventura, S.; Castro, C.; Hall, W. and Ng, M.H. (2002). Using genetic algorithms for data mining in web-based educational hypermedia systems. In *Proceedings of AH2002 workshop Adaptive Systems for Web-based Education,* Malaga, Spain.

44. Romero, C.; Ventura, S. and García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. In *Computers & Education, 51(1),* (pp. 368-384).

45. Romero, C.; Ventura, S.; Espejo, P.G.; Hervas, C. (2008) Data Mining Algorithms to Classify Students. In *Proceedings of the First International Conference on Educational Data Mining,* (pp. 8-17).

46. Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. In *Psychological Monographs: General and Applied, 80(1),* (pp. 1-26).

47. Schouten, B. and de Nooij, G. (2005). *Nonresponse adjustment using classification trees.* Discussion Paper 05001, Voorburg/Heerlen: Statistics Netherlands.

48. Scime, A. and Murray, G.R. (2007). Vote prediction by iterative domain knowledge and attribute elimination. In *International Journal of Business Intelligence and Data Mining, 2(2),* (pp. 160-176).

49. Simpson, O. (2004).The impact on retention of interventions to support distance learning students. In *Open Learning, 19(1),* (pp. 79-96).

50. Su, J.-M.; Tseng, S.-S.; Wang, W.; Weng, J.-F.; Yang, J.T.D. and Tsai, W.-N. (2006). Learning Portfolio Analysis and Mining for SCORM Compliant Environment. In *Educational Technology & Society, 9(1),* (pp. 262-275).

51. Superby, J.F. ; Vandamme, J.P. and Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Proceedings of the workshop on educational data mining, ITS'06,* (pp. 37–44).

52. Talavera, L. and Gaudioso, E. (2004). *Mining student data to characterize similar behavior groups in unstructured collaboration spaces.* Paper presented at Workshop on Artificial Intelligence in Computer Supported Collaborative Learning at European Conference on Artificial Intelligence. Retrieved July 29, 2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76.4034& rep=rep1&type=pdf

53. Vuk, M. and Curk, T. (2006). ROC curve, lift chart and calibration plot. In *Metodoloˇski zvezki, 3(1),* (pp. 89-108).

54. Wang, W.; Weng, J.; Su, J. and Tseng, S. (2004). *Learning portfolio analysis and mining in SCORM compliant environment.* Paper presented at the 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA. Retrieved July 29, 2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.5833&rep=rep1&type=pdf

55. Willging, P.A. and Johnson, S.D. (2004). Factors that influence students' decision to dropout of online courses. In *Journal of Asynchronous Learning Networks, 8(4),* (pp. 105–118).

56. Yukselturk, E. (2009). Do Entry Characteristics of Online Learners Affect Their Satisfaction? In *International Journal on E-Learning, 8(2),* (pp. 263-281).

57. Yukselturk, E. and Inan, F.A. (2006). Examining the Factors Affecting Student Dropout in an Online Certificate Program. In *Turkish Online Journal of Distance Education-TOJDE, 7(3),* Retrieved July 29, 2011, from http://tojde.anadolu.edu.tr/tojde23/pdf/article_6.pdf

58. Zang, W. and Lin, F. (2003). Investigation of web-based teaching and learning by boosting algorithms. In *Proceedings of IEEE International Conference on Information Technology: Research and Education, 2003,* (pp. 445-449).

59. Zhao, C. and Luan, J. (2006). Data mining: Going beyond traditional statistics. In *New Directions for Institutional Research, 131(2)*, (pp. 7-16).