Eğitim ve Bilim
2013, Cilt 38, Sayı 170

Education and Science
2013, Vol. 38, No 170

Corporate E-Learning Success Model Development by Using Data Mining Methodologies

Veri Madenciliği Yöntemleri ile Kurumsal E-Öğrenme Başarı Modeli Geliştirilmesi

Yasemin AYDOĞDU* Zuhal TANRIKULU**

Bogazici University

Abstract

The dynamic and more demanding nature of today's life conditions force people and corporations to invest in life-long education. It is important to make this continuous learning process more affordable and accessible to larger groups of people. At this point, e-learning seems to be more convenient way of learning than formal education especially for working adults because of their time and place constraints and their need for flexibility. The crucial concern is whether the e-learning process is useful or not and under what conditions it brings more value to adult learners. Thus, the core research question guiding this study is: What are the most significant factors influencing corporatee-learning success? The study aims to answer this question by developing e-learning success models via data mining. After a number of data preprocessing activities, a combination of descriptive and predictive data mining methodologies are applied on the data set. Most of the independent factors (learner demographics, learner experience, and course characteristics) are discovered to have power at different levels for explaining variance in e-learning success. Course program characteristics like content type, existence of certification are explored having a strong influence on the success of e-learning process.

Keywords: data mining, e-learning success factors, e-learning environment ile life-long learning

Öz

Talepkar ve dinamik yaşam koşulları, bireyleri ve kurumları yaşam boyu eğitime yatırım yapmaya zorlamaktadır. Önemli olan nokta, sürekli eğitimi mümkün olduğu kadar çok kişi için erişilebilir hale getirmektir. Elektronik öğrenme, zaman ve yer kısıtlarından dolayı özellikle çalışan yetişkinler için örgün eğitimden daha elverişli bir yöntem olarak görülebilir. Bu noktada önemli olan husus, sunulan elektronik öğrenme sürecinin yararlı olup olmadığı ve hangi koşullar altında öğrencilere daha fazla fayda sağlayacağıdır. Bu olgu, çalışmaya yön veren temel araştırma sorusunu ortaya çıkarmıştır: Kurumsal elektronik öğrenmenin verimliliğini ve başarısını etkileyen en önemli faktörler nelerdir? Çalışmada, veri madenciliği metotları kullanılarak geliştirilen elektronik öğrenme başarı modelleri ile bu sorunun cevaplanması amaçlanmıştır. Veri seti üzerinde yapılan bazı ön temizleme işlemleri sonrası veri seti üzerinde tanımlayıcı ve tahmine yönelik veri madenciliği modelleri uygulanmıştır. Bağımsız faktörlerin birçoğunun başarıdaki varyansı farklı seviyelerde açıklayabildiği sonucu çıkarılmıştır. Elektronik dersin türü, sertifikalı olup olmama gibi özelliklerin elektronik öğrenme başarısına daha güçlü etkisi olduğu görülmüştür.

Anahtar Sözcükler: Elektronik öğrenme başarı faktörleri, elektronik öğrenme ortamı, veri madenciliği, yaşam boyu öğrenme.

Introduction

Learning is an inevitable life-long process for human beings. The conditions in private life, work life, and social life force people to seek faster and less costly learning activities.

* Yasemin AYDOGDU, Bogazici University, Department of Management Information Systems, ya_yasemin@yahoo.com
** Assoc. Prof. Dr. Zuhal TANRIKULU, Bogazici University, Department of Management Information Systems, zuhal.tanrikulu@boun.edu.tr

Especially, in recent years, e-learning seems to be a very attractive alternative because of its potential advantages. Usage of the e-learning environment increases due to the advancements in technology. Computer technology makes the education process smoother with functions like the Internet and multimedia. These functions make learners and teachers more effective throughout the learning and teaching process. Audio-visual content makes online computer learning even more attractive (Akdağ & Tok, 2008). Moreover, enhanced motivation, coordination and teamwork, communication, improvement in critical thinking skills and opportunity for supplementary exercises are listed among outstanding benefits of e-learning (Arikan & Khezerlou, 2010). The integration of information technology with learning programs can be at different levels (micro, medium or macro level) based on the needs (Haşlaman, Mumcu, & Usluel, 2007). All in all, it seems that with the advantages of online learning stated above, it is possible to enrich the learning experience of human beings. However, as the demand for corporate online learning increases all around the world, effective e-learning process becomes more challenging due to diversified and dynamic environment.

E-Learning Around the World

Availability and accessibility of e-learning systems increases day by day with the advancements in technology. The technology enables required tools and methods to distribute educational content to the learners. In their study, Smith, Heindel and Torres-Ayala (2008) demonstrate that on-campus courses from various academic disciplines can be delivered online at an increasing rate. Their research at a major university in the South-eastern United States presents the result that the number of distance courses increased from 327 to 505 between 2002 and 2007. They also examined the distribution of online courses in those years. In the study, Biglan's taxonomy of academic disciplines is used which classifies the coursesbased on two dimensions as hard versus soft and pure versus applied. According to thet axonomy, courses are classified as hard-pure (HP), hard-applied (HA), soft-pure (SP) or soft-applied (SA). Mathematics, physics and, chemistry are regarded as hard-pure courses, while engineering and, applied mathematics are classified as hard-applied courses. On the other hand, social sciences, humanities and sociology are listed as soft-pure disciplines, whereas nursing, and education are considered among soft-applied disciplines. According to the study results, in 2002, soft-applied courses were 5% of the total, hard-pure courses were 8%, hard-applied courses were 38% and soft-applied courses comprised the largest proportion with 54% among all online courses. In 2007, the ratio of soft-pure courses increased to 15%, while soft-applied courses declined to 44%. The ratios of hard-applied and hard-pure courses also changed slightly with 35% and 6%, respectively. These results show that mostly content for applied sciences are delivered through an e-learning environment.

Another study mentions that an IT consulting and systems integration company has an e-learning environment which trains 1500 employees in 80 different countries and increased the number of online courses from 150 to 1500 in six years. While 80% of corporate training is online, only 20% is in the classroom (Gareiss, 2001).

The overall online enrollments are about 11% of all enrollments in 2008 and it is expected to increase to 20% by 2014. Importantly, even if the ratio is decreased from 69%, 55% of US degree-granting institutions are reported as not offering fully online courses. According to the report, 24% of adults who are over the age 25 take online programs and this ratio is predicted to increase to 35-40% by 2014 (Garrett, 2009).

The 2010 State of the Industry report on learning market in the USA by the ASTD (American Society for Training and Development) presents figures on e-learning market. It is stated that while formal learning hours which are available online were 23% of all formal learning hours in 2008, it increased to 27,7% in 2009. Moreover, average percent of learning hours which delivered via technology increased to 37% in 2009 (American Society for Training and Development [ASTD], 2010). ASTD's 2011 State of the Industry Report highlights that organizat ionsincreased

their spending per employee on learning and development by 13.5%. The report also states that there is still a growth in the use of technology in training, especially in terms of mobile learning. Furthermore, the report mentions that Fortune Global 500 companies set the goal of making 40.1% of formal learning hours be delivered via technology (American Society for Training and Development [ASTD], 2011).

Sybert and Lööf (2010) present the figures of the Internet usage for learning in European Union in 2010 based on Eurostat data. Among all Internet users, about 45% of users consult the Internet for learning, 32% searches information on courses and 7% participates in e-learning course. Eurostat data also provides Internet for learning statistics by education level, by age and by employment status. According to 2010 figures, users between 16 and 24 ages benefit from the Internet with highest percent. It is seen that as age increases, use of the Internet for learning decreases. While about 61% of people between 16 and 24 consult the Internet for learning, it is about 35% for people between 55 and 74. Similarly, the ratio decreases for e-learning course participation. While 9% of people between 16 and 24 participate in e-learning courses, only 3% of older people take e-learning courses. In their report, Sybert and Lööf (2010) make a comparison of the Internet for learning statistics between low, medium and high education levels. Based on the Eurostat data, it is presented that people with high education level take e-learning course with the highest ratio as 10%. Data on employment status shows that unemployed people benefit from the Internet for learning more than employed people.

E-Learning Success Factors

As the need and demand for online learning increases, improving the quality of learning and making education available for everyone becomes an important focus. Moreover, some financial concerns also make e-learning programs important. E-learning seems to be attractive by providing cost-effective educational systems as a result of technology usage. However, there are many dependent and independent factors influencing the outputs of corporate e-learning process. Asdiversity of learners with different demographics, backgrounds, business needs and capabilities increases, it is even more important to set a dynamic e-learning environment by taking all factors into consideration.

To develop effective e-learning systems and ultimately, to make use of its benefits and to increasesatisfaction of the learners, researchers have identified critical e-learning success factors which are summarized below.Study results on e-learning success factors can be reviewed in five major categories: instructor related, learner related, course related, e-learning system-related and other external factors.

1. The instructor has an effect on the e-learning success (Levy, 2006; Ozkan & Koseler, 2009; Paechter, Maier, & Macher, 2010; Selim, 2007). The instructor's attitude, teaching methods and guidance are considered as important factors for an effective e-learning process.
2. E-learning success also depends upon some learner-related factors. Learners' demographics (Levy, 2006; Mutlu, Yılmaz, & Erorta, 2004; Ong & Lai, 2004; Ozkan & Koseler, 2009; Robertsen, Trobridge, Leong, Mitra, & Hullett, 1997), IT skills (Kerr, M. S., Kerr, M. C., & Rynearson, 2006; Selim, 2007), time management skills (Kerr et al., 2006; Selim, 2007) and learning and academic skills (Kerr et al., 2006) influence the e-learning process. Online learning environment should be enriched by reusable and personalized content. The content can be customized based on the learner's profile via some technologies such as metadata tagging which tags the content by XML code and matches the content with learner's profile (Gareiss, 2001). XML coding is a base technology to create semantic web which helps to provide suitable and well-organized content based on learner profile and business demands. It enables the communication of human and machine on a semantic basis and makes the learner find and combine the most proper material based on his/her preferences (Stojanovic, Staab, & Studer, 2001). In

addition, frequency of e-learning usage and previous e-learning experience are important determinants for success or failure (Mutlu et al., 2004; Selim, 2007). Another key factor is the learner's attitude towards using technology (Robertsen et al., 1997). Moreover, the learner's perceived effectiveness (Ozkan&Koseler, 2009), motivation (Kerr, et al., 2006; Paechter et al., 2010) and achievement goals (Paechter et al., 2010) should be taken into consideration for better learning outcomes. Moreover, communication and collaboration among learners and instructors increases the quality of e-learning process (Levy, 2008; Paechter et al., 2010).

3. An e-learning course is a crucial part of the learning process. Course content, design and quality (Chao & Chen, 2009; Levy, 2006; Ozkan & Koseler, 2009; Paechter et al., 2010) should be taken into consideration to make learners experience a useful learning program. Enjoyment and usefulness of course program should not be ignored while designing course programs (Ozkan & Koseler, 2009).
4. E-learning system - related factors also are explored for improving the e-learning experience. Technology, system design and quality (Chao & Chen, 2009; Levy, 2006; Ozkan & Koseler, 2009; Selim, 2007) affect the learner's performance and the learning program's effectiveness. Some specific system related factors also are studied. Accessibility, performance, security and standard compliance are listed as important factors for evaluating e-learning system effectiveness (Kor & Tanrikulu, 2008). Usability, interaction, functionality, reusability, evaluation, appropriateness, design, interoperability, and accessibility of the system are presented as key factors in the study by Tanrikulu, Tugcu and Yilmaz, in 2010. Furthermore, availability for individual learning process (Chao & Chen, 2009; Paechter et al., 2010) and existence of synchronous learning (Chao & Chen, 2009) are proposed as important characteristics of the system. Importantly, the existence of a grading and monitoring system is a key factor for correct evaluations and feedback (Chao & Chen, 2009). Lastly, non-existence of technical support is considered a factor for e-learning failure (Selim, 2007).
5. Other external factors like ethical, legal and environmental issues and trends (Ozkan & Koseler, 2009) are presented as having influence on e-learning process.

Data Mining and E-Learning

The literature study above shows that there is an increasing demand fore-learning around the world. That is because especially multinational corporations need a flexible training environment in which they can train employees without time and place boundaries and by which they can easily offer a customized content based on employee profiles and business demands. E-learning meets those needs and is an attractive solution for corporations. In order to get return on investment and to establish an effective corporate e-learning environment, it is important to understand the most important e-learning success factors among the ones summarized above.

In order to analyze these critical success factors, data mining process is applied in this study. It aims retrieving useful knowledge from data by four major steps as data collection, preprocessing, applying data mining methods (descriptive or predictive), evaluating and interpreting results. It can be applied to very large data sets with various types of data. It can overcome high number of dimensionality, high number of classes or heterogeneousness of data (Maimon & Rokach, 2008). It is not possible to make database queries or to monitor a very huge and multidimensional data set in order to derive statistical results. It is aimed to understand similar groups in the data set, to measure the strongest influencers such as content, demographics on e-learning success and to establish corporate e-learning success models via proper data mining methodologies.

Problem

The dynamic and demanding nature of today's life conditions force human beings to invest continuously in their self-development. The need for life-long education is realized by a higher number of people and corporations. In recent years, other than in-classroom educational programs, establishment of online learning programs is considered a part of life-long education. To make use of an online learning environment effectively, the process of online learning should be explored and improved. People spend time with e-learning while also fulfilling daily responsibilities. Similarly, corporations spend time and budget to build required technical infrastructure and to restructure some organizational processes, accordingly. While both people and corporations are spending resources for this learning activity,it is crucial to understand whether the e-learning process really adds value. To get the return on all investments, the critical point here is to determine and to improve the factors related with e-learning success.

Research Questions

Core research question of this study focuses on identifying the factors influencing learner success. The total score of the learner on e-learning exams is represented by the term 'learner success'. Specific research questions are as follows:

A. What are the course program-related significant factors for learner success?

B. What is the effect of learners' previous e-learning experiences onlearner success?

C. What are the demographics-related significant factors for learner success?

Objectives

The goal of this study is to develop a learner success model and to understand the main contributors of e-learning systems. It aims to be able to give e-learning stakeholders useful directions and recommendations on how to treat to important e-learning related factors, and ultimately, on how to enhance e-learning systems further and make them much more contributive for the learner.

Hypotheses

To derive concrete results, hypotheses are produced related to the research questions. Statistical tests are applied and results are analyzed for each hypothesis.

Course program-related factors vs learner's success

- H1: There is a statistically significant association between learner's success and course program content which is either vocational or skill development. (Programs with vocational content aim to train the users in an area of specialization related to the banking sector such as fraud management, lending and budgeting. Programs with skill development content aim to contribute to the personal development of the users such as team work, communication and collaboration and stress management)
- H2: There is a statistically significant association between learner's success and duration of course program.
- H3: There is a statistically significant association between learner's success and whether course program is certified or not.

Learner's previous e-learning performance vs learner success

- H4: There is a statistically significant association between learner's success and number of online course programs previously taken by the learner.
- H5: There is a statistically significant association between learner's success and learner's

previous e-learning success (in average).

Learner's demographics vs learner's success

- H6: There is a statistically significant association between learner's success and learner's age (younger, middle age or older).
- H7: There is a statistically significant association between learner's success and learner's gender (male or female).
- H8: There is a statistically significant association between learner's success and learner's education level (high school, associate, bachelor or graduate degree).
- H9: There is a statistically significant association between learner's success and learner's functional occupation (marketing and sales, information technology (IT) or business support).
- H10: There is a statistically significant association between learner's success and learner's hiearchical occupation (operational, supporting, low level decision making or high level decision making).
- H11: There is a statistically significant association between learner's success and learner's geographical region (Marmara, Ege, Karadeniz, İç Anadolu, Doğu Anadolu or Güneydoğu Anadolu).
- H12: There is a statistically significant association between learner's success and learner's current work experience (low experienced (0-2 years), middle experienced (3-7 years) and high experienced (over 7 years).
- H13: There is a statistically significant association between learner's success and formal education success of the learner (unsuccessful, relatively successful or successful).

Methodology

Research Design

In this study, the process of knowledge discovery in databases is applied and data mining methods are used to explore the significant factors on e-learning success. Regression analysis and a decision tree are conducted to develop an e-learning success model. The learner's course score is selected as key indicator of effectiveness. The variance in learner score is measured based on learner related and course program-related factors. Age, gender, education, geographical region and, occupation in terms of functional group and hierarchical level and work experience are introduced as main demographics of learners. Learner average success and previous e-learning experience which is measured by the number of previous e-learning programs are used as indicators for learners' e-learning history. Course program (CP) - related characteristics such as course program content, duration and certification also are included in the study.

Sample

This research study is based on a database that consisted of actual data about online educational programs applied to the personnel of a private bank of Turkey. It includes data about learners, course programs, educations, exams and education evaluations. The database contains information about 5500 learners, 90 educations, 90 exams and 220 course programs. The data set contains about 45000 records before data preprocessing tasks. In the data set, data on vocational courses like budgeting, consumer loaning, capital market committee and skill development courses like social relationships, communication, and team-work are stored.

Sample Distribution and Reliability

The Kolmogorov-Smirnov test shows that there is no normal distribution in the data set ($p=0.000<0.05$), so that it is more meaningful and reliable to apply non-parametric tests. The reliability of the data set is measured by the ratio of Cronbach's alpha. It shows the level of inner consistency in the data set. Alpha is measured as 0.702 and is acceptable for applying statistical test, since it is greater than the threshold ratio $\alpha \geq 0.70$ (Nunnally, 1974).

Data Analyses

In this study, a knowledge discovery in databases (KDD) process is applied. Knowledge discovery and data mining process in e-learning aims to retrieve useful knowledge from data. It contains four major steps: Data collection, preprocessing, applying data mining methods and evaluating and interpreting results. Data mining models and tests are used in the study, since it doesn't seem possible to analyze huge amounts of data with informal monitoring or similar methods. Data mining enables exploration of hidden and important data patterns in huge data sets(Romero, Ventura,& Garcia, 2008). In this big multidimensional data set, there are learners with different demographics such as age, gender, education level and so on and these learners have different previous learning experiences. There are also courses with different content type, duration and so on. In this study, other than exploring new variables in the data set such as previous learning experience which is derived based on the previous courses, most importantly, it is analyzed what happens if different variable values exist in the same environment, how they affect e-learning success (score) all together and which characteristics gets the priority in determining e-learning score among all variables.

By the use of predictive clustering analysis, it is evaluated which combination of values results in high e-learning score. It is aimed to explore similarities based on existing characteristics in the data set and putting the similar samples into clusters. A good clustering has the ability to find all or some of the hidden patterns in the data set and produce high quality clusters with high intra-class similarity and low inter-class similarity (Ganti, Gehrke, & Ramakrishan, 1999; Gibson, Kleinberg,& Raghavan, 1998; Giudici, 2003; Wang H, Wang W., Yang, & Yu, 2002). By the use of predictive logistic regression model, it is measured that whether dependent variable-learner success depends on independent variables such as demographics and course characteristics. It is derived how strongly learner success is influenced by a unit change in each predictor or independent variable. The aim of regression models is to explore whether the dependent variable can be explained as a function of some independent variables or determinants (Giudici, 2003; Larose, 2006). By the use of decision tree as a predictive model, based on information provided like learner's age, occupation or course program content, a prediction is made on whether this learner will be successful or not. By the use of predictive models, it is possible to write 'if conditions' and to make predictions on e-learning score which is not possible via queries on a relational database. It seems that outcomes of data mining analysis are important for continuous improvement of e-learning systems. Not only academic leaders and administrators, but also learners and educators can make use of explored information as a feedback (Romero& Ventura, 2007).

Distinct data sets are generated from the database and SPSS 17.0 (Statistical Package for the Social Sciences) tool is used for applying data mining tests. K-means clustering analysis and statistical correlational tests as descriptive methods, logistic regression models and decision tree by CHAID (chi-squared automatic interaction detector) growing method as predictive methods are applied to the data set. Spearman's rank correlation coefficient, Kruskal-Wallis and Wilcoxon Mann-Whitney U tests are used for correlational analysis.

Results

In this section, results of preliminary analysis for similar group discovery, results of correlational analysis and e-learning success prediction models are presented.

Preliminary analysis for similar group discovery: cluster analysis

To explore the hidden patterns in the data set with respect to learner success, a k-means clustering analysis is conducted by iterate and classify method. Two clusters are discovered in the data set as can be seen in Table 1 below. Cluster 1 is the cluster of successful programs with a total score mean of 0.96. Cluster 2 has an average score of 0.67 and can be considered as unsuccessful cluster.

The major dissimilarities between two clusters are the characteristics of the programs. Course program duration is much longer in successful cluster. Furthermore, 99% of the programs collected in the successful group are certified. Lastly, course program content differentiates the two groups. The unsuccessful group took only vocational course programs, whereas there are mostly skill development programs in the successful group.

Table 1.

Final Cluster Centers

		Cluster1	Cluster2
C_Total_point	Learner 's average program score	,96	,67
C_Age	Learner's age	,62	,57
Male	Gender	,48	,38
Female	Gender	,52	,62
E_HighSchool	Learners with high school degree	,30	,23
E_Associate	Learners with associate degree	,14	,14
E_Bachelor	Learners with bachelor degree	,55	,61
E_Grad	Learners with graduate degree	,01	,02
C_WorkExperience	Learner's average work experience	,26	,20
R_Marmara	Learner's from Marmara region	,31	,34
R_IcAnadolu	Learner's from İç Anadolu regioN	,30	,28
R_Ege	Learner's from Ege region	,10	,12
R_Karadeniz	Learner's from Karadeniz region	,11	,11
R_DoguAnadolu	Learner's from Dogu Anadolu region	,06	,05
R_GüneydoguAnadolu	Learner's from Güneydogu Anadolu region	,11	,10
HO_HighLvl	Learners in the high level of organizational hiearchy	,44	,32
HO_LowLvl	Learners in the low level of organizational hiearchy	,01	,00
HO_Supporting	Learners in the supporting level of organizational hiearchy	,10	,12
HO_Operational	Learners in the opeartional level of organizational hiearchy	,46	,55
FO_MarketingSelling	Learners in the marketing and selling function	,76	,77
FO_IT	Learners in the information technology function	,12	,11
FO_CorpBusinessSupport	Learners in the corporate business support function	,12	,13
C_CPDuration	Duration of online course program	,45	,03
C_CPCompDuration	Learner's course program completion duration	,15	,01
C_NoofCP	Learner's number of previous course programs	,49	,44
C_AvgCompDuration	Learner's average course program completion duration	,15	,11
C_UserAvgSuccess	Learner's average success in onlien courses	,83	,77
C_NotCertificated	Is the program certificated? (No)	,01	1,00
C_Certificated	Is the program certificated? (Yes)	,99	,00
C_CPVocational	Vocational course program	,49	1,00
C_CPSkill	Skill development course program	,51	,00

Correlations with learner success

To explore the relationships and to discover significant factors for learner success, correlational analyses are conducted for three dimensions -course program characteristics, learners' previous e-learning performance and learner demographics. As mentioned by Neideen and Brasel (2007), there is a significant relationship, if p value is equal or less than 0.05. Based on this statistical condition, significant correlations are determined.

Firstly, course program characteristics are analyzed. According to the correlational tests by Spearman's rho, there is a significant association between learner's success and course program's content (CP_Content), course program's certification (CP_Certificated), course program's duration (CP_Duration) with correlation coefficients presented in Table 2 below. H1, H2 and H3 are accepted.

Table 2.

Correlations – Course Characteristics

			CP_Content	CP_Certification	CP_Duration
Spearman's Rho	Learner Success	Correlation Coefficient	,491**	,691**	,497**
		Sig. (2-tailed)	,000	,000	,000
		N	18963	18963	18963

Secondly, influence of learners' previous e-learning performance is analyzed. According to the correlational tests by Spearman's rho, there is a significant association between learner's success and learner's total number of previous course programs (Numberof_CPs) and learner's previous e-learning success with correlation coefficients presented in the Table 3 below. H4 and H5 are accepted.

Table 3.

Correlations – Previous E-Learning Experience

			LearnerAvgSuccess	NumberofCPs
Spearman's Rho	Learner Success	Correlation Coefficient	,360**	,121**
		Sig. (2-tailed)	,000	,000
		N	18963	18963

Lastly, significant correlations between learners' demographics and success are explored. According to the Spearman's rho correlational test statistic, there is a significant association between learner's success and learner's age, gender, education level, functional occupation, hierarchical occupation, and current work experience with correlation coefficients presented in Table 4 below. H6, H7, H8, H9, H10 and H12 are accepted.

Sperman's rho correlational test results show that there is no significant correlation between learner's e-learning success and region with p=0.191>0.05) and between learner's e-learning success and formal education success (branch score) with p=0.466>0.05. H11 and H13 are rejected.

Table 4.

Correlations - Learner Demographics

			Age	Gender	Education	Occ Functional	Occ Hierarchical	Region	Curr Work Exp	Branch Score
Spearman's Rho	Learner Success	Correlation Coefficient	,215**	,047**	-,063**	-,026**	,112**	-,010	,141**	,006
		Sig.2 tailed	,000	,000	,000	,000	,000	,191	,000	,466
		N	18963	18963	18963	18963	18963	18963	18963	14043

After significant correlations are explored, further tests are applied to prove the difference among groups. By applying Kruskal Wallis tests, groups which are different from each other in terms of learner success are uncovered.

Sample Kruskal Wallis test statistics presented in Table 5 below show that at least two groups that participated in a different number of e-learning programs do not have identical distribution in terms of success. In Table 6 below, a sample Kruskal Wallis test statistic is alsopresented for age groups. There are younger (between 0-29), middle age (29-35) and older (>35) age groups. Asymp. significance level which is less than 0.05 results in rejection of null hypothesis which claims that all populations have identical distributions. When the significance level for age is examined in Table 6 below, it is proven that at least two age groups perform differently in online course programs and have different success levels. Similarly, Table 7 below shows that success level changes based on course program content (vocational content or skill development).

Table 5.
Kruskal Wallis Test Statistics-NoP

	U_TotalPoint (Binned)
Chi-Square	282,347
df	2
Asymp. Sig.	,000

Grouping Variable: Number of Programs

Table 6.
Kruskal Wallis Test Statistics-Age

	U_TotalPoint (Binned)
Chi-Square	1127,696
df	2
Asymp. Sig.	,000

Grouping Variable: Age

Table 7.
Kruskal Wallis Test Statistics-CP

	U_ TotalPoint (Binned)
Chi-Square	4564,186
df	1
Asymp. Sig.	,000

Grouping Variable: Course Program Content

Table 8 below shows the mean score for each age group and Mann Whitney- 2 independent samples test shows that there is statistically significant difference between younger - older age groups and middle - older age groups.

Table 8.

Mean Score for Age Groups

Age	Mean	N	Std. Deviation
Younger	70,21	5984	21,610
Middle age	70,46	5432	20,053
Older	81,11	7547	20,726
Total	74,62	18963	21,478

Multinomial logistic regression model for learner success

The objective is to model learner success as a function of independent variables (predictors), so that it is possible to explore the relationships as a whole and to establish a prediction model based on factors. Based on the information about a new case, the model enables distinguishing different groups (Giudici, 2003; Larose, 2006). In this study, the model differentiates successful learners from unsuccessful learners.

The dependent variable is learner success (total score) in the model. Region and formal education success variables which are not significant according to correlation results are excluded from the model. Included independent variables are as follows:

- Course content (vocational or skill development content)
- Course duration

- Course program completion duration
- Course certification
- Previous number of course programs
- Learner average success
- Age
- Gender
- Occupation (functional and hierarchical)
- Education
- Work experience

Nagelkerke's pseudo r-square statistic is 67% which satisfies the threshold level (>65%) and indicates the reliability of this multinomial logistic regression model. As a result of the model, predictors which have significant influence in explaining the variance in learner success are selected for the function of learner success (factors with sigma<0.05). Estimate values are the coefficients of predictors in multinomial logistic regression equation and shows the strength of influence, which is the change on the dependent variable (learner success) due to a unit of change in predictors (Giudici, 2003; Larose, 2006). According to model results, course program characteristics such as course content, course certification, course program duration; previous e-learning performance indicators as number of previous e-learning programs, learner's average success and demographic as age, gender, education, functional occupation are reported as significant predictors. Course program certification can be proposed as the strongest predictor with the highest estimate value, whereas work experience and hierarchical occupation are insignificant in differentiating learner's success group. The equation is: P(unsuccessful/successful)=6,136*Certification(No) +4,914*Course program content

1. (Vocational course) -0,862*Course duration (short)- 1,444*Course duration (medium)

-0,519*Course completion duration (short) +0,245* Course completion duration (medium)

-0,565*Number of e-learning programs(Less)

-0,237 * Number of e-learning programs (Medium) -0,137*Learner average success

-0,037*Age -0,553* Education (High school) -0,147*Gender(Female)

-1,138*Functional Occupation(Marketing&Sales) -1,205*Functional Occupation(IT)

The model is also generated by including region and formal education success which are insignificant according to correlational results. It is seen that these variables are not selected as significant and most of the parameter estimates slightly differ from the previous modeland results can be seen in equation (2) as follows:

2. P(unsuccessful/successful)= 6,043*Certification(No) +4,653*Course program content (Vocational course) -1,603*Course duration (short) - 2,057*Course duration (medium)

-1,197*Course completion duration (short) +0,264* Course completion duration (medium)

-0,577*Number of e-learning programs(Less)

-0,241 *Number of e-learning programs(Medium) -0,140*Learner average success

-0,038*Age -0,135*Gender(Female) -1,031*Functional Occupation(Marketing&Sales)

-1,096*Functional Occupation(IT)

Classification accuracy. It is stated that classification accuracy should be at a proper level to be able to accept the model as useful and accurate. The estimate of by chance accuracy criteria is calculated and compared with the overall accuracy rate of the model. An accurate model is defined as providing at least 25% improvement in the estimate by chance accurate rate (Hair,

Anderson, Tatham, & Black, 1998).

The overall classification accuracy of modelin equation (1) is 69,7% as presented in Table 9 below and is greater than the estimate of by chance accuracy criteria which is calculated as 41,6 % based on the information in Table 10 below.

Table 9.
Case Processing Summary

		N	Marginal Percentage
Learner Success	Unsuccessful	6337	33,4%
	Medium	6367	33,6%
Successful		6259	33,0%

Table 10.
Classification Accuracy

	Predicted			
Observed	Unsuccessful	Medium	Successful	Percent Correct
Unsuccessful	3920	2263	154	61,9%
Medium	1966	4313	88	67,7%
Successful	245	1032	4982	79,6%
Overall Percentage	32,3%	40,1%	27,5%	69,7%

A decision tree model for learner success

Decision tree is another method of predictive data mining. Based on information provided like learner's age, occupation or course program content, decision tree classification model generates a prediction on whether this learner will be successful or not.

For this decision tree model, the CHAID growing method is used at a significance level of 0.05. For validation reasons, a split sample validation method is applied and the sample is divided equally for training and testing. Misclassification costs remain the same for incorrect prediction of each group (successful and unsuccessful). Variables used in regression model are included in decision tree model and insignificant variables (region and formal education success) are excluded from the model. The model is trained for prediction of two groups: Successful and unsuccessful.

Classification accuracy. Overall classification accuracy is 87.7% and the details can be seen in Table11 below:

Table 11.

Classification Accuracy

		Predicted		
Sample	Observed	Unsuccessful	Successful	Pereent Correct
Training	Unsuccessful	3148	81	97,5%
	Successful	691	2492	78,3%
	Overall Percenttage	59,9%	40,1%	88,0%
Test	Unsuccessful	3027	81	97,4%
	Successful	679	2397	77,9%
	Overall Percentage	59,9%	40,1%	87,7%

The decision tree is generated based on the included independent factors and given data set. Similar to the regression model explained above, course program content, certification and duration are selected in the model while predicting learner success. Additionally, learners' total number of e-learning programs as a previous e-learning performance indicator is used for classification. The model also takes age and hierarchical occupation information among demographics of the learner as stronger predictors.Education and functional occupation are not

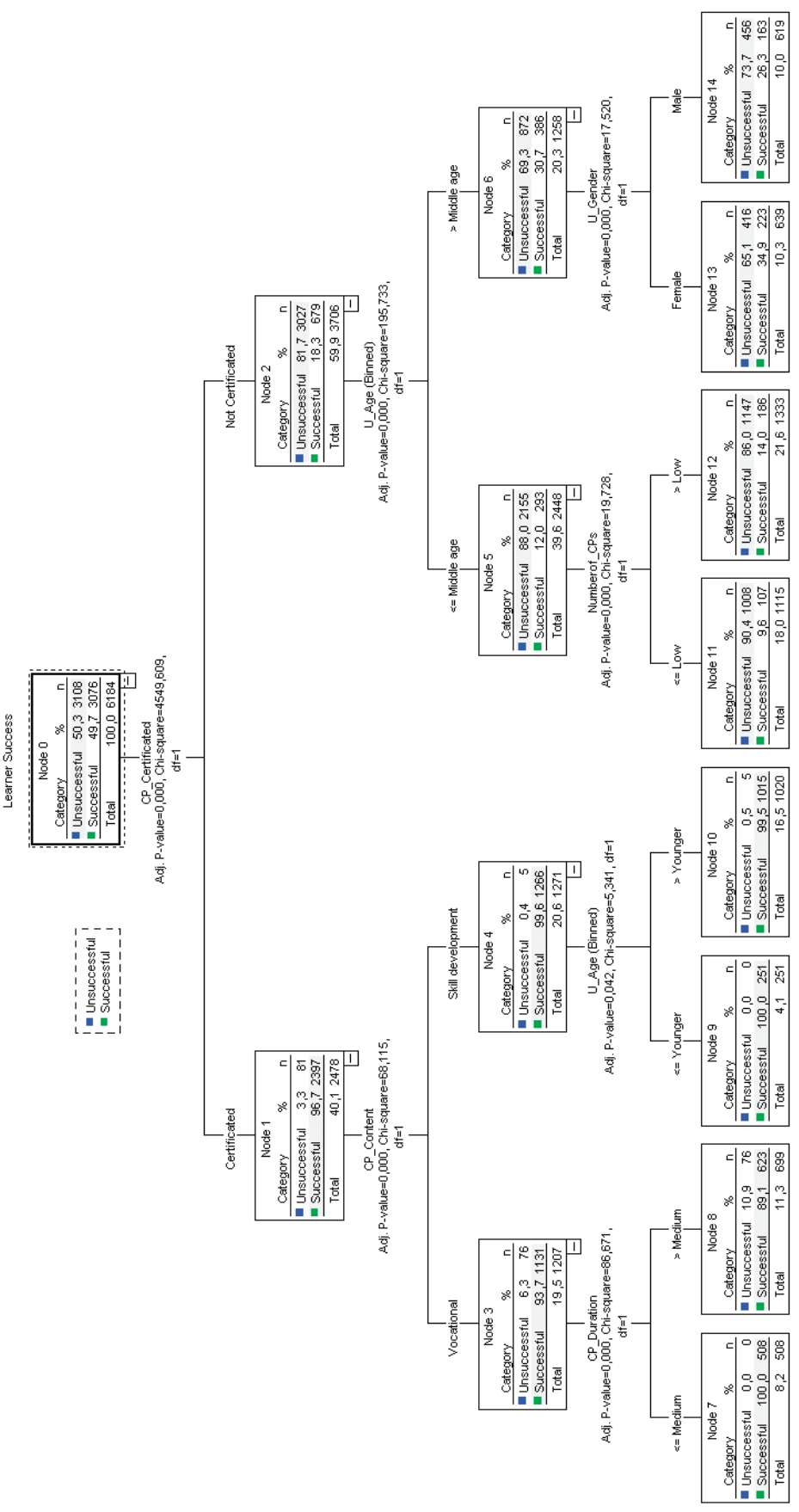

Figure 1. Decision tree classification-learner success

chosen as significant predictors as the same in the regression model.

Terminal nodes of decision tree are node 7 through node 14. A new incidence is classified as successful or unsuccessful based on association rules generated. For example, the information about a new incidence is a learner who is older, has a bachelor degree and has high participation in e-learning programs. A current course program titled "Fraud Management in Banking" is a certified vocational course program which is assigned a really longer duration to be completed. From this information, the tree model firstly checks for whether the course program is certified or not. According to the information, this is a certificated course program. Before classifying this incidence, the model is interested in course program content as a result of the logic. Since the course program is a vocational oriented course program, the next factor to be checked is course program duration. Longer duration information is given in this new incidence, so that the final node is node 8 which predicts that learner will be successful by a probability of 89.1%. Association rule followed for this case is as follows:

IF (CPCer = "Certificated") AND (CPCont = "Vocational") AND

(CP_Duration IS MISSING OR (CP_Duration > "Medium")) THEN

Node = 8

Prediction = 3

Probability = 0.898072

Discussion

In this study, in addition to correlational tests, two predictive models - regression analysis and decision tree classification are conducted in order to establish success models. The reason is, to make a comparison between these two separate methods and to see whether there are differences in the results of different methods. Both models support the correlational analysis results. Course program related factors are selected as major predictors in both analysis.Among all the factors influencing e-learning success, the strongest and the weakest influencers are discovered as course program certification and functional occupation, respectively.

When the results compared with the situation in formal education, it is seen that some course program-related factors are explored to be significant also in formal classical education. Content familiarity is presented as a success factor (Schönwetter, Clifton, & Perry, 2002). Additionally, Vermunt (2005) states that different academic disciplines effect learning patterns. Depending upon the discipline, learners show the characteristics of meaning, reproduction, application directed or undirected learning. Furthermore, even if the correlation is not very strong, a relationship is discovered between demographics and e-learning success. One question is whether this outcome is in alignment with the one in formal classical education. In classical education, learning success is explored to be influenced by learner characteristics, as well (Beekhoven, De Jong, & Von Hout, 2003; Schönwetter et al., 2002). Gender (Beekhoven et al., 2003; Schönwetter et al., 2002), locus of control, test anxiety, high school GPA (Schönwetter et al., 2002), repeating the course, subjective chance of success (Beekhoven et al., 2003) are among key influencers of success.

According to the correlational tests by Spearman's rho, course program characteristics are strong influencers and there is a significant association between learner's success and course program's certification (CP_Certificated), course program's duration (CP_Duration) and course program's content (CP_Content) with correlation coefficients 0.691, 0.497 and 0.491, respectively. These strong correlations make it important to focus on course program specifications. The decision tree model provides further information on the strongest influencer - course certification. It is seen that while about 97% of learners are successful in certificated programs, only about 18% are successful in non-certificated programs. According to the decision tree, if it's a certificated program, the strongest predictor of success is course content type (soft skill or vocational program). Deeper investigation may be performed to understand the effect of different content type. As mentioned, capability and background of each learner differs which at the end affects

how much the learner can understand the content and gain useful information.

For this reason, different learner profiles should be taken into consideration and customized content solutions should be provided by the systems. At this point, artificial intelligience systems seem to be an effective solution. They can make connections between the learner and course program content based on the learner's profile and can make recommendations to the learner to increase e-learning efficiency (Chen, 2008; İnceoglu, Uğur, & Aslan, 2004; Romero & Ventura, 2007). Furthermore XML coding and semantic web also seem to be promising technologies for content customization (Gareiss, 2001; Stojanovic et al., 2001).

Learners' motivation level is stated as a success factor in e-learning (Kerr, et al., 2006; Paechter et al., 2010). In this study, certification is assumed as a motivating or triggering factor for the learners and the study results supports the fact that learners are more enthusiastic when they are provided some motivating factors such as certification. As shown in Table 2, learner success (score) has the strongest correlation with certification (whether the program offers certification or not) with a correlation coefficient as 0.691. It is obvious that learners should participate in the program voluntarily in order to absorb useful information from the course as much as possible. In this study, the effect of certification is obvious. It can be regarded as a kind of motivator which encourages learners to spend time in an e-learning program effectively. Similar motivating factors like reward systems or reflecting positive effect on their annual scorecards based on their e-learning outcomes can be designed.

Conclusion

This study provides statistical results on e-learning effectiveness to provide useful insights for practioners, designers and decision makers of e-learning systems. Generated success models have the ability to classify learners into groups with respect to success.

All the analysis results prove that while most of the learner demographics are weakly correlated with learner success (score). Among learner demographics, age with 0.215 correlation coefficient has the strongest relationship with learner success. However, there isa strong correlation between learner success and course program specifications. Results demonstrate that learners are more successful in certified programs. Certification has a significant correlation with learner score with a coefficient of 0.691. Course content is also explored among key influencers in the generated success models. It has a correlation coefficient of 0.491 which proves a significant correlation with learner score. Moreover, course duration has a strong relationship with learner score with a correlation coefficient of 0.497. Other than course program characteristics, learners' previous e-learning experience has an effect on learner score. It is determined that as the number of online learning programs participated by the learner increases, then, the learner performs better. According to the study results, previous e-learning experience (number of previous courses) is strongly correlated with learner score with a coefficient of 0.121. These correlational results are supported by regression and decision tree models. Regression model equation indicates that change in course certification has the strongest influence on the dependent variable learner success. Similarly, the course certification variable is discovered to be the premier predictor in the learner success decision tree model.

All these results may provide guidance for selecting the improvements areas especially in corporate e-learning education. It should be highlighted that the effectiveness or ineffectiveness of e-learning is absolutely a result of nested relationships from demographics to course program specific attributes.

This study is conducted based on large data sets. However, there are some limitations related to data which prevents analysis to some extent. Most importantly, there is no normal distribution in the data set. As a result, parametric tests can not be conducted. Many academical resources indicate that the number of non-parametric tests is limited but they are almost as powerful as parametric tests.

Acknowledgement

This research was realized at the Information Systems Research and Application Center of Boğaziçi University. Authors would like to thank enocta® for their support.

References

Akdağ, M., & Tok, H. (2008). The Effects of Traditional Instruction and PowerPoint Presentation–Supported Instruction on Student's Achievement. *Education and Science*, 33 (147), 26-34.

Arikan, A., & Khezerlou, A. (2010). Prospective English language teachers' views on computer and paper-based instructional materials in developing language components. *Procedia Social and Behavioral Sciences*, 2(2), 4006–4009.

ASTD American Society of Training and Development (2010).The 2010 State of the Industry Report, Research report published by ASTD.

ASTD American Society of Training and Development (2011).The 2011 State of the Industry Report, Research report published by ASTD.

Beekhoven, S., De Jong, U., & Von Hout, H. (2003). Different courses, different students, same results? An examination of differences in study progress of students in different courses. *Higher Education*, 46(1), 37-59.

Chao, R., & Chen, Y. (2009). Evaluation of the criteria and effectiveness of distance e-learning with consistent fuzzy preference relations. *Expert Systems with Applications*, 36(7), 10657–10662.

Chen, C.-M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers&Education*, 51(2), 787–814.

Ganti, V., Gehrke, J.,&Ramakrishan, R. (1999). CACTUS clustering categorical data using summaries. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 73-83), San Diego, USA.

Gareiss, D. (2001). E-Learning around the world. *Information Week*, 4(9), 63-64.

Garrett, R. (2009). Online higher education market update 2008 New York and national data. [Online] Retrieved January 30, 2011, from http://www.slideshare.net/alexandrapickett/richard-garretts-online-higher-education-market-update-2008-national-new-york-data.

Gibson, D., Kleinberg, J.,&Raghavan, P. (1998). Clustering categorical data: An approach based on dynamic systems. *Proceedings of the 24th VLDB Conference*, (pp. 311-322), New York, USA.

Giudici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry.* New York : J. Wiley.

Hair, J., Anderson, R., Tatham, R.,&Black W. (1998). *Multivariate Data Analysis*, 5th. Edition, Prentice-Hall Int., Inc.

Haşlaman, T., Mumcu, F. K.,&Usluel, Y. K. (2007). The integration of information and communication technologies in learning and teaching process: A lesson plan example. *Education and Science*, 32(146), 54-63.

İnceoglu, M. M., Uğur. A.,&Aslan, B. G. (2004). "Intelligent Approaches for Web-based E-Learning Systems." *Proceedings of the first International Conference on Innovations in Learning for the Future: e-Learning*, (pp. 244-250), Istanbul, Turkey.

Kerr, M. S., Rynearson, K.,&Kerr, M. C. (2006). Student characteristics for online learning success. *Internet and Higher Education*, 9(2), 91–105.

Kor, B.,&Tanrikulu, Z. (2008). Evaluation of E-Learning Systems: Standards&Choosing Test Tools. *Proceedings of 2nd International Future-Learning Conference on Innovations in Learning for the Future 2008: e-Learning*, (pp. 764-771), Istanbul, Turkey.

Larose, D. T. (2006). *Data Mining Methods&Models*. Hoboken, NJ : Wiley Interscience.

Levy, Y. (2006). *Assessing the value of e-learning systems*. Hershey, PA: Information Science Pub.

Levy, Y. (2008). An empirical development of critical value factors (CVF) of online learning activities:An application of course program theory&cognitive value theory. *Computers&Education*, 51(4), 1664–1675.

Maimon, O.,& Rocah, L. (2008). *Soft computing for knowledge discovery and data mining*. Boston, MA: Springer Science+Business Media, Inc.

Mutlu, M. E., Yılmaz, U.,&Erorta, Ö. Ö. (2004). "Efficiency of E-Learning in Open Education."*Proceedings of the first International Conference on Innovations in Learning for the Future: e-Learning*, (pp. 361-372), Istanbul, Turkey.

Neideen, T.,&Brasel, K., (2007). Understanding Statistical Tests. *Journal of Surgical Education*, 64(2), 93-96.

Nunnally, J. C. (1975). Psychometric theory- 25 years ago & now. *Educational Researcher*, 4(10), 7-14 + 19-21.

Ong, C-S.,&Lai, J-Y. (2004). Gender differences in perceptions&relationships among dominants of e-learning acceptance. *Computers in Human Behavior*, 22(5), 816–829.

Ozkan, S.,&Koseler, R. (2009). Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation. *Computers&Education*, 53(4), 1285–1296.

Paechter, M., Maier, B.,&Macher, D. (2010). Students' expectations of,&experiences in e-learning: Their relation to learning achievements&course satisfaction. *Computers&Education*, 54(1), 222–229.

Robertsen B., Trobridge G., Leong J.-A., Mitra A.,&Hullet, C. R.(1997). Toward Evaluating Computer Aided Instruction: Attitudes, Demographics, Context. *Evaluation&Program Planning*, 20(4), 379-391.

Romero, C.,&Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135–146.

Romero, C., Ventura, S.,&Garcia, E. (2008). Data mining in course management systems: Moodle case study&tutorial. *Computers&Education*, 51(1), 368–384.

Selim, H. M. (2007). Critical success factors for e-learning acceptance: Confirmatory factor models. *Computers&Education*, 49(2), 396-413.

Schönwetter, D. J., Clifton, R. A.,&Perry, R. P. (2002). Content familiarity: Differential Impact of Effective Teaching on Student Achievement Outcomes. *Research in Higher Education*, 43 (6), 625-655.

Smith, G., Heindel, J. A.,&Torres-Ayala, A. T. (2008). E-learning commodity or community: Disciplinary differences between online courses. *Internet&Higher Education*, 11(3-4), 152–159.

Stojanovic, J., Staab, S.,&Studer, R. (2001). Elearning based on the semantic web. Paper presented` at the World Conference on the Web&Internet, October 23-27, 2003, Orlando, Florida, USA. [Online]Retrieved Jully 5, 2012, from http://lufgi9.informatik.rwth-aachen.de/lehre/ws02/semDidDesign/lit/SemanticWebELearning.pdf

Sybert, H., & Lööf, A. (2010). *Industry, trade & services*. European Commission Eurostat Report (Data in Focus 50/2010). [Online] Retrieved February 03, 2011, from http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-QA-10-050/EN/KS-QA-10-050-EN.PDF.

Tanrikulu, Z., Tugcu, C.,&Yilmaz, S. (2010). e-University: Critical Success Factors. *Procedia – Social and Behavioral Sciences*, 2(2), 1253-1259.

Vermunt, J. D. (2005). Relations between student learning patterns&personal&contextual factors and academic performance. *Higher Education*. 49(3), 205-234.

Wang, H., Wang, W., Yang, J.,&Yu, P.S. (2002). Clustering by pattern similarity in large data sets. *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data,* (pp. 394-405), Madison, USA.