Proc. of the Intl. Conference on Advances in Information Processing and Communication Technology - IPCT 2016.

ISBN: 978-1-63248-099-6 doi: 10.15224/ 978-1-63248-099-6-37

Institutional Effectiveness Prediction Using Data Mining Techniques

Blerta Abazi - Caushi, Florije Ismaili, Agron Caushi, Xhemal Zenuni, Bujar Raufi

***Abstract*— The development of Information Technology has generated large amount of data in various areas. Organizations are deploying different analytic techniques to evaluate rich data sources in order to extract useful information within the data and utilize this in further decision making. In this paper the different approaches to IS development as well as how the investment in technology and IS contribute to increase the student headcount are investigated. Moreover, a considerable amount of work is done in classifying and building models for university classification according to the investment of the institutions in IT services and their risk management. Examinations about educational technology services with emphasis in e-learning technologies are done as well.**

***Keywords*— Educational Data mining, information extraction, decision making.**

I. Introduction

The development of Information Technology has generated large amount of data in various areas. Organizations are deploying different analytic techniques to evaluate rich data sources in order to extract useful information within the data and utilize this in further decision making. More recently, higher education has been strongly influenced by global trends, especially as a result of the call by governments for universities worldwide to improve their performance and efficiency [4]. Rising quality in competitive education environments have forced universities to adopt new strategies in order to improve their performance. Consequently, the higher education sector has turned to Enterprise Resource Planning (ERP) systems in the hope to deal with the changing environment.

Blerta abazi - Caushi
South East European University
Macedonia

Florije Ismaili
South East European University
Macedonia

Agron Caushi
South East European University
Macedonia

Xhemal Zenuni
South East European University
Macedonia

Bujar Raufi
South East European University
Macedonia

This is the reason that researchers and developers from educational community started exploring the potential adoption of similar techniques for gaining insight into online learner's activities.

Educational data mining is emerging as a research area using different research approaches for converting educational data into useful information in order to improve institutional effectiveness [4]. The purpose of this research is analyzing

- The IT environments and different approaches to IS development.
- The impact of information systems in decision making process and the functioning of Higher Education Institutions.

II. Related Work

Recently an important work is done towards the usage of data mining techniques in Education. An evaluation of a comprehensive literature review of relevant researches done in the field of educational data mining is provided by paper [1] and [2].

In the first literature review, the researches done in the area are classified into five areas: a) Survey of papers published in Educational Data Mining, b) Predicting Academic Performance with Pre/Post Enrollment Factors, c) Comparison of Data Mining Techniques in predicting academic performance, d) Correlation among Pre/Post Enrollment Factors and Employability and e) Other areas of Education.

The second literature review provides background for understanding current knowledge on Learning Analytics (LA) and Educational Data Mining (EDM) by qualifying the research don based on four discrete stages: a) searching the literature – data collection, b) reviewing and assessing the search results – selection of primary studies, c) analyzing, coding and synthesizing the results, and d) reporting the review.

Another survey in educational data mining is done by paper [3] which focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness.

After analyzing the presented literature we conclude that the research done in the field of educational data mining is concentrated on faculty and student evaluation, student behavior modeling and prediction of performance, determining student's success or satisfaction for a particular course, course registration planning, predicting the enrollment headcount, the importance of employers on soft-skills and academic reputation, educational resources handling etc. in other words the research is focused more in

Proc. of the Intl. Conference on Advances in Information Processing and Communication Technology - IPCT 2016.

ISBN: 978-1-63248-099-6 doi: 10.15224/ 978-1-63248-099-6-37

improving institutional effectiveness with applying data mining techniques in improving student learning processes.

A literature review of ERP systems in HEI is done by [4], however, unlike other applications little research has been conducted regarding these systems in a university environment. Most existing evaluation studies of ERPs focus on technical issues or implementation processes, these do not provide an explanation about ERPs effects, or if ERPs work well or poorly with a specific user in a particular setting.

In contrast from the current work presented above our research is focused in important IT issues related to universities information systems such as IT service delivery and the use of educational technology services in teaching and learning process.

III. Data Mining: The Research Base

Data mining (DM), is the process of transformation of large amounts of data into meaningful patterns and rules [8]. Both expert opinion and data mining techniques play an important role at each step of procedures used to examine and transform data.

Further, data mining can be divided in two types:

- Directed: to predict a particular data which involves finding unknown values/relationships/patterns from known values point.
- Undirected: to create groups of data or find patterns in existing data.

Data mining combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases. Its primary goal is to extract knowledge from data to support the decision-making process [9]. A data mining process generally includes the following four steps illustrated in figure 1.

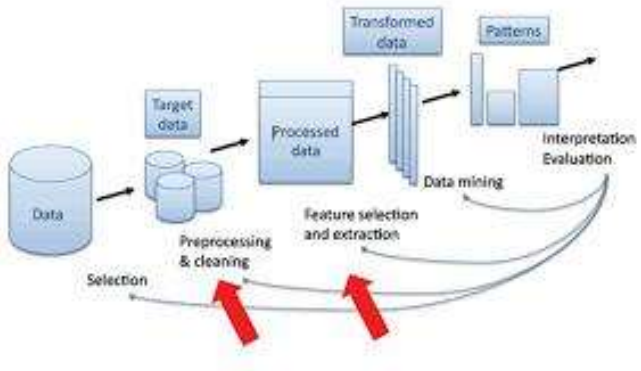

Figure1. Main steps of the data mining process [11]

Data acquisition: The first step is to select the types of data to be used.

Preprocessing data: Once the data is selected, it is then preprocessed for cleaning and transforming to improve the effectiveness of discovery.

Model Building: This step refers to decision on the type of DM operation; selecting the DM technique; choosing the DM algorithm; and mining the data.

A. *Techniques Used in Data Mining*

As outlined above, the data mining endeavor involves many steps. Furthermore, these steps require technologies from other fields. In particular, methods and ideas from machine learning, statistics, database systems, data warehousing, high performance computing, and visualization all have important roles to play [10]. There are several categories of data mining problems for the purpose of prediction and/or for description:

- Classification: developing a model that maps a data item into one of several predefined classes.
- Clustering: segmenting a dataset into clusters, each of which shares common and interesting properties.
- Regression: building a model that maps data items into a real-valued prediction variable.

B. *Educational Data Mining*

Higher education is data rich but information poor. Although universities collect data about students, academic activities, administrative and operational functions, still many institutions struggle with how to transform those data into useful information [4, 5].

Educational Data Mining (EDM) is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. The discipline focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness [6].

C. *Educational Data Mining Motivation*

This research is based on EDUCAUSE database which has large amount of data having several attributes stored in different databases. The data is associated with queries, responses and other intradepartmental issues. Therefore the data mining can convert those data into extraction of new patterns and information which ultimately is for University benefit and growth.

Questions to response

- What are the different approaches to IS development?
  - Vendor based,
  - Homegrown,
  - Open source.
- What is the impact of Information systems in the functioning of HEI?
- Does the investment in technology and IS contribute to increase the student headcount?

Proc. of the Intl. Conference on Advances in Information Processing and Communication Technology - IPCT 2016.

ISBN: 978-1-63248-099-6 doi: 10.15224/ 978-1-63248-099-6-37

IV. Highlights of ERP Implementations

A large number of Universities have implemented or are planning to implement an ERP system. It is important to understand why institutions are investing on ERP and what institutions plan to do next. In order to do this, module 3 which includes questions about educational technology service functions and facilities provided by central IT and other units is analyzed. Topics include: student technology, faculty instructional technology/LMS support, classroom and learning space support, multimedia services and distance education services.

At the study's inception, we assume that most institutions rely primarily on vendor based information systems, which proves to be the case.

Among 959 universities, 549 responded the question about the product that was operational for the primary undergraduate admissions system at institutions. 13, 66 % of universities uses homegrown solutions, while 86, 34% of them uses other type of system.

TABLE I. UNDERGRADUATE ADMMISSIONS

	Number of universities
Homegrown	75
Other	474
Total	549

Of the 474 universities 461 have used single product while 13 universities have used more than one product. Among them, 455 universities have used vendor products while 19 of them have used open source products.

TABLE II.OTHER UNDERGRADUATE ADMMISSIONS PRODUCTS

	Number of universities
Vendor product	455
Open source	19
Single product	461
More than 1	13
Total	549

Figure 2 illustrates the distribution of respondents by used technology platform type for undergraduate admissions:

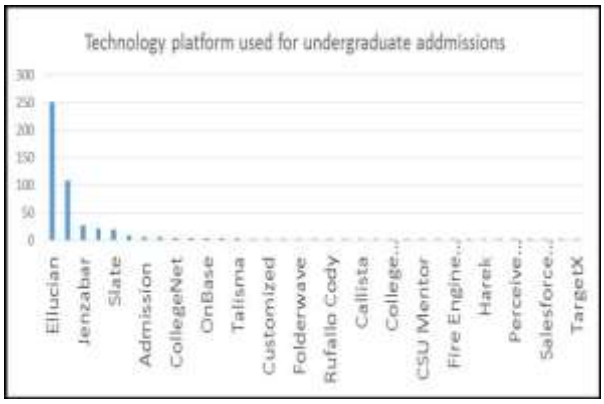

Figure 2. Technology Platform used for Undergraduate Admissions

The results from survey respondent indicates that for all core information systems only 9.85% uses homegrown solutions while 90.15% uses other type product including vendor based and open source products. According to the results, Ellucian products are adopted in a large number of universities for advancement/fundraising, CRM, data warehouse, financial aid, financial management, human resources, procurement, grants management and student information systems followed by Oracle PeopleSoft, Blackbaud Raiser's Edge and SAP. In contrast, Google Apps and Microsoft's products are the preferred products for staff/student e-mail systems, whereas learning (course) management system Blackboard Learn and Moodle are at the top of the list. The provided results proves the hypothesis that most institutions rely primarily on vendor based information systems.

In important issue to evaluate at this point is the investment of the institutions in IT services and their risk management. The survey is focused in the following concerns:

A. Investment in IT services is adequate to meet institutional needs

B. We have enough qualified staff devoted to IT risk management

C. We have an adequate budget devoted to IT risk management

TABLE I. THE INVESTMENT OF INSTITUTIONS IN IT SERVICES AND THEIR RISK MANAGEMENT RESULTS

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
A	81	280	224	221	37
B	156	365	202	105	22
C	163	339	214	104	21

Among all surveyed institutions about 41 % didn't respond to the questions related to investment of the

Proc. of the Intl. Conference on Advances in Information Processing and Communication Technology - IPCT 2016.

ISBN: 978-1-63248-099-6 doi: 10.15224/ 978-1-63248-099-6-37

institutions in IT services and the state of IT risk management. The result of the respondent institutions is shown in Table III.

Based on the response results, for every option, after invoking J48 to perform a data mining session the universities are classified in three main classes as Universities that invests in IT services as "YES", Universities that do not invests in IT services as "NO", neutral Universities as "Neutral" and Universities that didn't respond are classified as "Unknown".

The analysis uses a dataset holding information about individuals (Universities) who were classified as "YES", "NO", "Neutral" and "Unknown" using all attributes mentioned above. We want to decide on a best set of attributes defining the classes contained in the data. Stated another way, we wish to test the possibility of building an accurate supervised learner model.

The screen containing the resultant decision tree follows

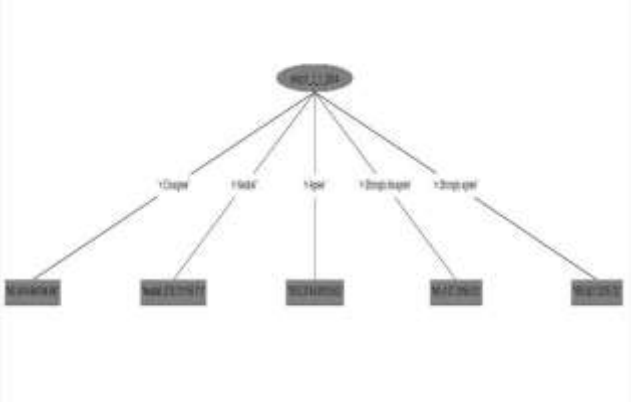

Figure 3. Decission tree for university classification based on their IT investment and risk management.

In the resulting tree, the attribute representing the first question represents the top-level node of the decision tree. Upon scrolling through the tree, we observe a classification accuracy of 97, 62% which ensures us that the model will classify Universities in appropriate manner.

A. *Educational Technology Services Positioning*

Using E-learning tools and technologies in online training is another very important issue in university education process. Therefore, this section includes examinations about educational technology services with emphasis in e-learning technologies.

Our goal is to identify how public and private universities deploy e-learning, e-portfolio, apply information of literacy requirements and deploy interactive learning.

The results are obtained using different clustering techniques. The result received by clustering based in Simple kMeans where atributes are the number of universities using e-learning and other attributes related to learning technologies. The results shows that universities are classified as 515 (58%) public and 370 (42%) private universities that deploy e-learning technologies.

Clustering by Simple k-Means algorithm by 4 clusters shows that 313 public universities deploy e-learning, are experimenting with e-portfolio, deploy hybrid courses, do not plan to apply information of literacy requirements and deploy interactive learning. In the same way we interpret the other clusters. 211 public universities deploy e-learning, are considering to deploy e-portfolios, deploy hybrid courses, do not plan to apply information of literacy requirements and are experimenting with interactive learning. The third cluster includes 184 public universities which deploy e-learning, are experimenting with e-portfolio, deploy hybrid courses, are considering to try information of literacy requirements and are experimenting with interactive learning. The fourth cluster includes 177 private universities which deploy e-learning, are considering applying e-portfolio, deploying hybrid courses, are not planning to try information of literacy requirements and deploy interactive learning.

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall
                   1          0          1          1
                   1          0          1          1
                   1          0          1          1
                   1          0          1          1
Weighted Avg.      1          0          1          1

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 627   0   0   0 |   a = 1
   0 126   0   0 |   b = 2
   0   0  51   0 |   c = 3
   0   0   0  70 |   d = 4
```

Figure 4. SMO Clustering results of Universities according to deployment of e-technologies

In figure 4 are shown the results with SMO (Support Vector Machine). We have 4 clusters, classified according to the e-learning class where class a=1 shows that 627 universities already deploy e-learning technology, class b=2 shows that 126 universities are experimenting with it, class c=3 shows that 51 universities are considering to apply this technology and class d=4 shows that 70 universities are still not planning to apply e-learning. The precision and recall in this case are both 1.

The association rule with apriori algorithm is applied to the same dataset. The results indicates that from 115 universities that apply interactive learning and they all deploy hybrid courses and information literacy requirements with a confidence of 0.97, 111 of them apply e-learning. The universities that intend to deploy blogs, deploy broadly hybrid courses and interactive learning are 104 and 100 apply e-learning with 0.96 confidence.

B. *The Impact of Investment in Technology and IS*

Regression models all fit the same general pattern. There are a number of independent variables, which, when taken together, produce a result — a dependent variable. The regression model is then used to predict the result of an unknown dependent variable, given the values of the

Proc. of the Intl. Conference on Advances in Information Processing and Communication Technology - IPCT 2016.

ISBN: 978-1-63248-099-6 doi: 10.15224/ 978-1-63248-099-6-37

independent variables [7]. Linear regression will be used to analyze how the investment in technology and IS contribute to increase of the student headcount?

The model will identify how the total amounts central IT received during the prior fiscal year depends from funding categories such as operating appropriation, capital appropriation, appropriation from revenue generated from student IT fee, Revenue from sale of services, student headcount, student FTE etc.. Figure 5. Shows the results of regression model.

According to the regression model, it is obvious that Student FTE have a great impact on the total amount that central IT receives, while in opposite correlation with student head count. The result indicates that the total amount depends directly from the students that have enrolled and follow the courses.

```
Linear Regression Model

m1q16a_total_2015 =

     113.7475 * Student_FTE_2013 +
     -67.3946 * Student_head_2013 +
       0.9008 * m1q16a_10_2015 +
       0.9979 * m1q16a_1_2015 +
       0.8253 * m1q16a_2_2015 +
       0.9464 * m1q16a_3_2015 +
       0.9887 * m1q16a_4_2015 +
       1.0268 * m1q16a_5_2015 +
       1.0243 * m1q16a_6_2015 +
       1.3146 * m1q16a_7_2015 +
       0.9764 * m1q16a_8_2015 +
       1.0529 * m1q16a_9_2015 +
-587061.0797

Time taken to build model: 0.09 seconds
```

Figure 5. Regression model for total ammount received by Central IT

v. Conclusion

One of the most recent challenge that higher education faces today is finding patterns from IT services data in order to improve institutional effectiveness. In this paper the different approaches to IS development as well as how the investment in technology and IS contribute to increase the student headcount are investigated. Moreover, a considerable amount of work is done in classifying and building models for university classification according to the investment of the institutions in IT services and their risk management. Examinations about educational technology services with emphasis in e-learning technologies are done as well. As educational data mining emerges which have great impact to the university functioning, there is an opportunity to enforce the methods listed above to accomplish a variety of goals. Furthermore, the presented methods can be utilized to study new hypotheses and answer new research questions.

References

[1] P. Thakar, A. Mehta, and Manisha, “Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue,” International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 15, January 2015 .

[2] Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. Educational Technology & Society, 17 (4), 49–64.

[3] IR Huebner, “A survey of educational data-mining research,” in Research in Higher Education Journal, v19 Apr 2013

[4] Berland, M., Baker, R.S. & Blikstein, P. (2014). Educational Data Mining and Learning Analytics: Applications to Constructionist Research. Technology, Knowledge and Learning, Volume 19, Issue 1-2, pp 205-220.

[5] M. Bienkowski, M. Feng, B. Means, “Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics”, Center for Technology in Learning SRI International , October 2012 .

[6] G. Siemens and R. Baker. “Learning analytics and educational data mining: towards communication and collaboration.” In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), Simon Buckingham Shum, Dragan Gasevic, and Rebecca Ferguson (Eds.). ACM, New York, NY, USA, 252-254. 2012.

[7] IBM, Developers Work, Data mining with WEKA, Part 1: Introduction and regression, 2015.

[8] O. A. Nassar and N. A. Al Saiyd, "The integrating between web usage mining and data mining techniques," Computer Science and Information Technology (CSIT), 2013 5th International Conference on, Amman, 2013, pp. 243-247.

[9] Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 2011.

[10] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques (Third Edition), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 2011.

[11] Eko Indarto, Data Mining, Accessed 2016; http://recommender-systems.readthedocs.io/en/latest/datamining.html