Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining

Fadzilah Siraj and Mansour Ali Abdoulha

College of Arts and Sciences
University Utara Malaysia
Sintok, Kedah, Malaysia
fad173@uum.edu.my

Abstract

To date, higher educational organizations are placed in a very high competitive environment. To remain competitive, one approach is to tackle the student and administration challenges through the analysis and presentation of data, or data mining. This study presents the results of applying data mining to enrollment data of Sebha University in Libya. The results can be used as a guideline or roadmap to identify which part of the processes can be enhanced through data mining technology and how the technology could improve the conventional processes by getting advantages of it. Two main approaches were used in this study, namely the descriptive and predictive approaches. Cluster analysis was performed to group the data into clusters based on its similarities. For predictive analysis, three techniques have been used Neural Network, Logistic regression and the Decision Tree. The study shows that Neural Network obtains the highest results accuracy among the three techniques.

Keywords: Data Mining, Education, Enrollment, Neural Network, Logistic Regression

1. Introduction

To date, higher educational organizations are placed in a very high competitive environment and are aiming to get more competitive advantages over the other business competitors. These organizations should improve the quality of their services and satisfy their customers (industry, government). To remain competitiveness among educational field, these organizations need deep and enough knowledge for a better assessment, evaluation, planning, and decision-making. The majority of the required knowledge that stored in the educational organization's database can be extracted from the historical and operational data. Therefore, one approach to effectively tackle the student and administration challenges is through the analysis and presentation of data, or data mining (DM).

DM helps organizations to use their current reporting capabilities to discover and identify the hidden patterns in databases. The extracted patterns are then used to build data mining models, and hence can be used to predict performance and behaviour with high accuracy. As a result of this insight, universities are able to allocate resources more effectively. DM may, for example, give a university the information necessary to take action before students quit their study, or to efficiently assign resources with an accurate estimate of how many male or female will register in a particular program ([1]).

University has collected large amounts of student data for years; however this data is typically not put in a form of improving its use. To date, universities are data-rich but information poor. Many of them did not take the advantage of DM in analyzing and uncovering the hidden information within the student enrolment data. An attempt to uncover the hidden information will inevitably useful to produce knowledge that in effect improves management decision-making.

This study addresses usage and usefulness of DM and its applications on higher education databases particularly for understanding undergraduate's student enrolment data at Sebha University in Libya. It utilizes descriptive and predictive data mining approach in order to discover hidden information. Cluster analysis was performed to group the data into clusters based on its similarities. The clusters were also used as targets for prediction experiment. For predictive analysis, three techniques have been used Neural Network, Logistic regression and the Decision Tree. The study shows that Neural Network obtains the highest results accuracy among the three techniques.

978-0-7695-3648-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AMS.2009.117

2. Related Works

DM is defined as a term used in describing knowledge in databases. It is seen as a process of extracting and identifying useful information and subsequent knowledge from databases using statistical, mathematical, artificial intelligence and machine learning technique ([2]). DM applies modern statistical and computational technologies in its quest to expose useful patterns hidden within the large databases. It has proved itself as a powerful tool, capable of providing highly targeted information to support decision-making and forecasting for scientific, physiological, sociological, the military and business decision making. The predictive power of DM comes from its unique design – it combines techniques from machine learning, pattern recognition, and statistics to automatically extract concepts, and to determine the interrelations and patterns of interest from large databases ([3]).

It is highly necessary to determine the techniques of DM that are applicable in higher education environment. In fact, there are many algorithms that are similar in concept to stored procedures of object-oriented programming in that they are universally applicable. Almost all algorithms or models currently used in the business sectors are directly usable for research in higher education, especially in institutional researches except for Link Analysis which is only used in telecommunication companies to understand groupings associated with starting points ([4]). Furthermore, prediction from DM offers the college an opportunity to act before a student drops out or to plan for resource allocation with confidence gained from having complete records of all students reflecting their tracks of activities. Through data mining, a university could, for example, predict with 85 percent accuracy which students will or will not graduate. The university could use this information to plan required academic assistance on those students projected to experience such graduating difficulties.

The university's data can be used to inform solutions to a wide range of educational challenges. [5] listed group to explore differences, exploring growth over time, evaluating programs, and to identify the root causes of problems in education as one of the many ways data can be used. A study by [6] revealed that data is a strong predictor of the efficiency in the activities of school teams. The use of data is not only increased efficiency but also, to serve as a mediator for the positive effect of other factors. [7] considered the use of data as a central component of its business model to increase the achievement of the set objectives.

The data can also have a positive effect on people involved in the educational process. [8] observed that frequent usage of data in schools has metamorphosized into a more professional culture. Educators in their study have become greater collaborators during data/ decision-making process, and school business consequently has become a less "privatized" one. [9] noted that school leaders are involved in the use of data often develop a mindset of being responsible for their own destiny, increasingly able to find and use information to inform the school improvement. [10] noticed that the use of data has helped in raising expectations of teachers on their students, no positive changes in teachers' attitudes regarding the potential success of previously low-performing students.

The applications of DM in education sector is one of the most challenging tasks, this notwithstanding, its ability to offer a unique educational decision-making process is a good justification for the required stress involved. With the introduction of DM concept, decision makers (management) in the educational sectors will definitely find their jobs easier.

3. Methodology

The CRISP-DM methodology suggested by [11] is utilized in this study. This methodology involves six phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment as shown in Fig. 1.

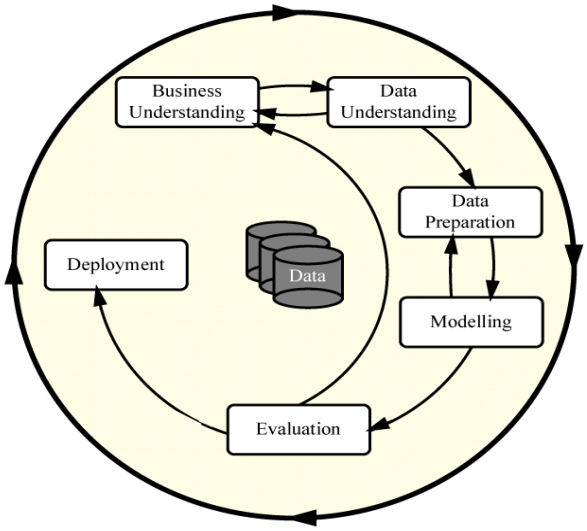
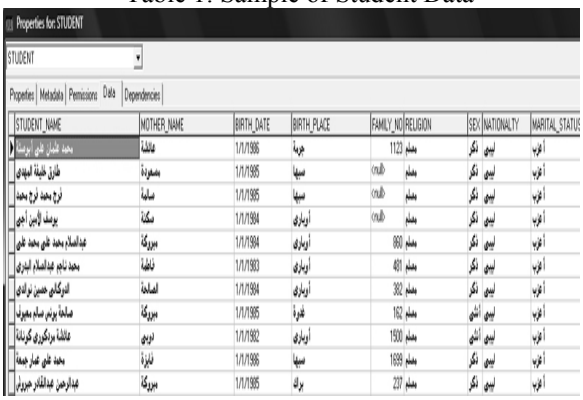


Figure 1: Steps of CRISP-DM Methodology (Adopted from Chapman *et al.,* 2000).

In this study, as the possible areas of tests depend on the data available, and the detailed business objective cannot be identified until the data is studied, consequently, this phase has to be performed in parallel with the data understanding and data preparation phase. The initial phase of Data Understanding focuses on understanding the study objectives and requirements from the student registrar office. The data understanding phase starts with an initial data collection and proceeds with actions in order to get familiar with the data.

The enrolment dataset contains 8510 instance from 1998 to 2006. An original student main table includes 38 attributes, 8 attributes are numerical and the others are of type categorical. Part of the original data is shown in Table 1.

Table 1: Sample of Student Data



As a result of preprocessing phase, the total number of data is reduced to 6830. Two types of data mining approaches were conducted in this study. The first approach is descriptive which is concerned on the nature of the dataset like the frequency table and the relationship among the attributes obtained using cross tabulation analysis (contingency tables). In addition clustering analysis is conducted to determine the similarity between the attributes of the dataset, and predictive data mining by using several prediction models (Decision Tree, Regression, Neural Network), Comparison between these models also conducted to determine the best model for the dataset.

4. RESULTS

Sebha University has been established in the year 1983. To date, Sebha University has several branches, they are located at Sebha, Ghat, Tragen, Brak, Morzoq, and Obari cities. Based on the information of all the faculties of Sebha University population distribution as shown in Fig 2, Dentistry (Sebha), Sport (Ghat), Arts (Tragen) and Science (Tragen) has the small population ranging from 1% to 4% with number of students less than 350 over 8510 of all university population. This indicates that the university should put this fact in consideration in coming years to know why the students at such locations are low. However if the faculties are grouped by the cities, Sebha faculties (Science, Arts, Medicine, Dentistry, Law and Agronomy) have the ratio of 55% of university population.

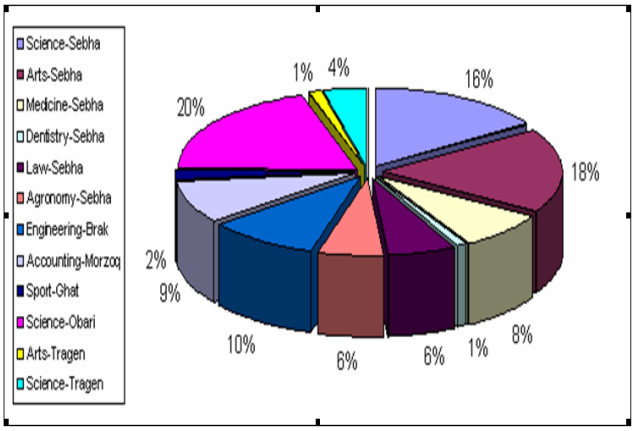


Figure 2: Distribution Student Population

The descriptive statistics, particularly cross tabulation analysis was carried out to discover the relationship between the attributes (Fig. 3).

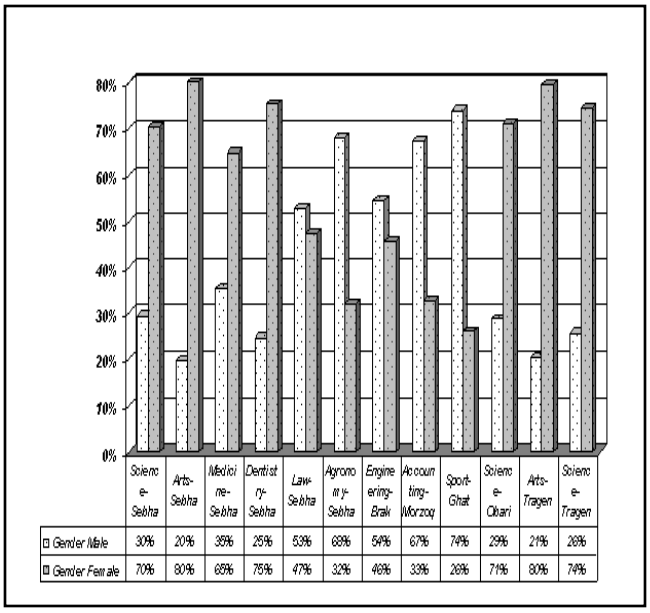


Figure 3: Faculty with Respect to Gender

Based on the results shown in Fig. 3, majority of the registered students are female with ratio of 58% in all university population and almost 42% are male. Program such as Science (Sebha, Obari and Tragen), Arts (Sebha and Tragen), Medicine (Sebha) and Dentistry (Sebha) are more popular to female students, with ratio of 80% in some faculties, and ranging from 65% to 80%. In contrast, other degree programs such as Law (Sebha), Agronomy (Sebha), Engineering (Brak), Accounting (Morzoq) and Sport (Ghat) have more male students than female ranging from 50% to 75%.

Further analysis was carried out to determine the relationship between faculty, gender and student status. The student status was classified into ***Enroll, Move, Expel, Quit*** and ***Completed the Study***. From the analysis, it is observed that higher percentage of female students ***Completed the Study*** compared to male students undertaking Science (Obari and Sebha), Arts (Sebha) and Medicine (Sebha). On the other hand, higher percentage of male students undertaking Sport (Ghat) and Law (Sebha) completed their studies as shown in Fig. 4.

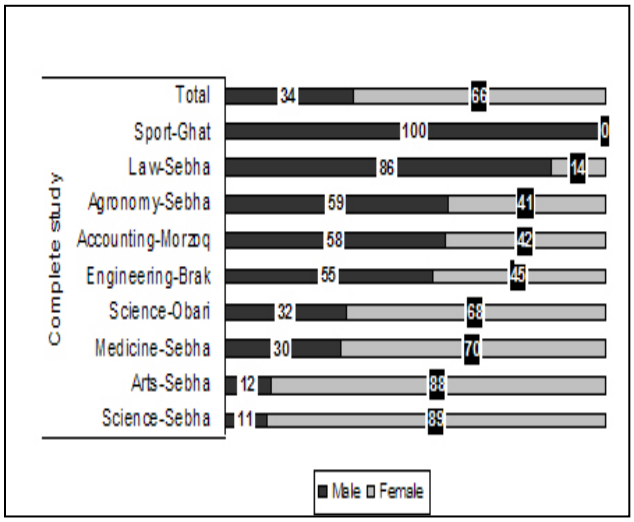


Figure 4: Faculty with Respect to Student Status (*Comp. Study*) and Gender

Based on results exhibited in Fig. 5(a) and 5(b), more male compared to female students have been expelled from the university. When Gender is cross tabulated with Student Status, most of the students who quitted Arts program (Sebha) are male (Fig. 5(a)). In addition, Agronomy (Sebha) program is not preferred by female students. Fig. 5(b) indicates students that have been expelled from continuing Engineering (Brak) and Science (Sebha) programs are male students. More than 50% of male students have also been expelled from continuing Arts (Sebha) degree program.

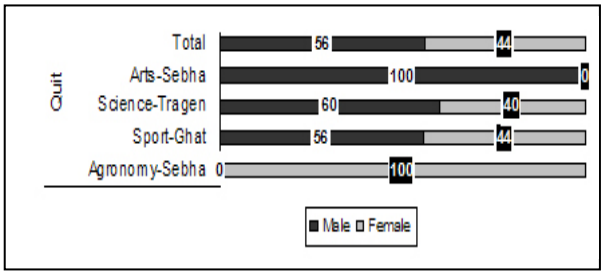


Figure 5(a): Faculty with Respect to Student Status ***(Quit)*** and Gender

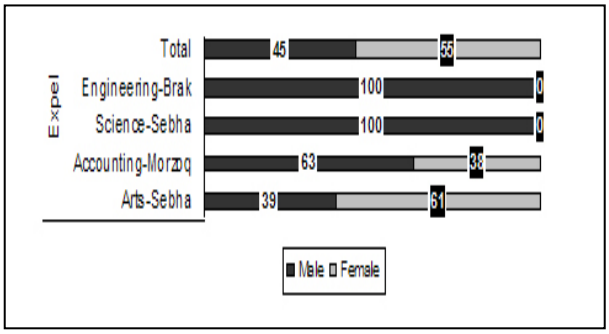


Figure 5(b): Faculty with Respect to Student Status ***(Expel)*** and Gender

Having performed cross tabulation analysis, the clustering network using Kohonen network has been carried out. As a result, 3 clusters has been identified (Cluster 0, 1 and 2). Further investigation was performed to carry out to determine the relationship between variables such as ***Religion, Gender, Nationality, State, Degree Owned, Faculty, Student Status, Admission Type, Housing Status*** and ***Register Type*** with Clusters (Table 2).

Table 2: The Correlation between enrollment attributes and ***Clusters***

	Degree Owned	***Faculty***	***Housing Status***	***Nationality***
Correlation Coefficient	**.156(**)**	**-.760(**)**	**-.287(**)**	**.332(**)**
Sig. (2-tailed)	**.000**	**.000**	**.000**	**.000**

*	Correlation is significant at the 0.05 level (2-tailed).
**	Correlation is significant at the 0.01 level (2-tailed).

Clearly, the correlation between variables ***Cluster*** and ***Faculty*** is significantly strong ($p = 0.00$, $r = -0.760$) while between ***Cluster*** and ***Nationality*** is

medium (p = 0.00, r = 0.332). The overall result of determining the characteristic of each cluster and comparison between all clusters is shown in Table 3. The results exhibited in the Table 3 indicate that the faculties in cluster 0 and cluster 1 are different from cluster 2. Students undertaking Arts (Tragen) fall into Cluster 0, whereas the students at Sebha is clustered to Cluster 1 or Cluster 2. Programs such as Law (Sebha), Engineering (Brak) in university residence compared to those living outside campus (66% versus 34%). However, students in Cluster 0 are through government process (90%). Accounting (Morzoq), Sport (Ghat) and Argonomy (Sebha) are found in Cluster 0 or 1. In Cluster 0, higher percentage of students living in University Residence compared to those living outside the university campus (66% versus 34%). As for the faculty with respect to gender and cluster, higher percentage of female students compared to male in Cluster 1 (74% versus 26%). These female students are undertaking programs such as Science (Sebha and Obari), Arts, Medicine and Dentistry at Sebha. This also implies that females students prefer to undertake programs at Sebha. Further observation on the results also indicate that students are non-residence and admission through university selection process.

Table 3: Clusters Characteristic With Respect to Predictor's Variables

Variables	*Cluster 0*		*Cluster 1*		*Cluster 2*	
FACULTY	**Degree**	**Place**	**Degree**	**Place**	**Degree**	**Place**
	Science	Tragen	Science	Sebha	Science	Sebha
			Science	Obari	**Science**	**Obari**
					Science	**Tragen**
	Arts	**Tragen**	Arts	Sebha	**Arts**	**Sebha**
			Medicine	Sebha	**Medicine**	**Sebha**
			Dentistry	Sebha	**Dentistry**	**Sebha**
	Law	**Sebha**			**Law**	**Sebha**
	Engineering	Brak			Engineering	Brak
	Accounting	Morzoq			Accounting	Morzoq
	Sport	Ghat			Sport	Ghat
	Agronomy	Sebha			Agronomy	Sebha
GENDER	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
	52%	48%	26%	74%	44%	56%
HOUSING STATUS	**University Residence**	**Non-residence**	**University Residence**	**Non-residence**	**University Residence**	**Non-residence**
	66%	34%	41%	59%	41%	59%
ADMISSION CANDIDATOR FOR STUDENTS	Government	University	Government	University	Government	University
	90%	10%	5%	95%	2%	98%

For predictive analysis, three techniques have been used, namely the logistic regression, the decision tree and neural networks. For regression analysis, only independent variables **faculty** and **nationality** are significant to the regression prediction model with accuracy of 99.44%. In addition, these variables also have strong significant correlation with the dependent variable (Cluster). Decision tree analysis was performed by partitioning the data into training (70%), validation (15%) and testing (15%). Like regression analysis results, **Faculty** and **Nationality** are two important variables in decision analysis with respect to **Cluster**. Similar partitioning of data has been applied to Neural Network and the results show that the accuracy using Neural Network is 99.98 percent (versus logistic regression is 99.44 percent and the Decision Tree is 99.77 percent). Fig. 6 illustrates the lift chart for the three prediction models based on clustering results as the target.

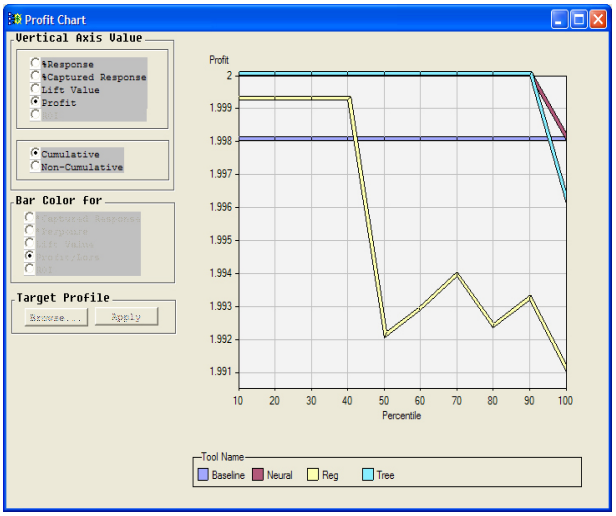


Figure 6: The Comparison Accuracy Between 3 Predicton Techniques

The lift chart also indicates that between 10-90% percentiles, both neural network and decision tree obtained the same accuracy. However, between 90-100% percentiles, neural network degrades slowly compared to decision tree. Hence, neural network is the better model among the three.

5. Conclusion

The descriptive statistics, particularly cross tabulation analysis presents a lot of useful information about the university data. In addition, it has been concluded that more female (63%) enrolled at the university compared to male (37%). In fact, female students tend to undertake Arts (Sebha and Tragen), Science (Tragen, Obari and Sebha) Dentistry (Sebha) and Medicine (Sebha). On the other hand, more male students tend to undertake several programs such as Sport (Ghat), Agronomy (Sebha) and Accounting (Morzoq). This may be due the fact that Sport (Ghat) and Accounting (Morzoq) is located in low population

area (Ghat and Morzoq). Furthermore, Agronomy (Sebha) is far from the city of Sebha, around 15 KM. Descriptive statistics and correlation analysis defined two a ttributes as the most importantly attributes, they are faculties and nationalities with respect to clusters, those attributes can significantly affect the student enrolment data among all other attributes.

The analysis conducted on the students that have been expelled from the university indicates that more male students are being expelled from the university compared to female students. In fact, 100% of the students that have been expelled from Engineering-Brak and Science-Sebha are male students. This matter is rather serious since the ratio of male to female total enrolment is about 1: 3. If this matter is not considered seriously by the university, this could lead to shortage of male students graduated with Science and Engineering degree.

Cluster Analysis was performed to group the data into clusters based on its similarities. In effect, the cluster results are used also as targets for prediction experiment. For predictive analysis, three techniques have been used: they are Neural Network (NN), Logistic Regression (LR) and the Decision Tree. The accuracy achieved more than 99% for Neural Network, Regression and Decision Tree. When further analysis was performed on the cluster, it is interesting to note that the cluster is able to distinguish between Libyan and non-Libyan students. In addition, some rule with regard to faculty and nationality can also be extracted. Hence, the prediction models based on clusters have shown significant result in exploring hidden information with Sebha University enrolment dataset.

The results of this study could be useful for those associated with the registration and education process of students in Sebha University in general, and in the registrar office in particular. Moreover, the results could assist registration planners to formulate proper and suitable plans for the university. The results will also help planners to revise for example the criteria for admission to the various student qualifications. Furthermore, the rules extracted from this study can help registrar office and university administrator to organize or restructure in order to plan necessary enhancement and improvement for enrollment purposes.

To improve the model, more attributes such as students year/semester of study and the academic achievement could be included to deliver other prediction models. In addition, it is recommended that the information and the delivered knowledge should be automated. The results obtained from this study also indicate to Sebha University in particular and all public universities in Libya as a whole to improve their proportion of students intake based on gender.

6. References

[1] Luan, J. (2004). Data Mining and Knowledge Management in higher Education Potential Application. *Proceedings of Air Forum, Toronto, Canada.*

[2] Efraim, T., Jay, E.A.,Tin-Peng, L. Ramesh, S. (2007). *Decision Support and Business Intelligent Systems* (Right Edition), Pearson Education, Inc.

[3] Edelstein, H. (1997). Data mining: Exploring the hidden trends in your data. *DB2 Online Magazine*. Retrieved from http:.//www.db2mag.com.

[4] Luan. J. (2001). *Data Mining as Driven by Knowledge Management in Higher Education-Persistence Clustering and Prediction*. Keynote for SPSS Public Conference, UCSF.

[5] Seifert, J. W. (2004). *Data mining: An overview*. Congressional Research Service, the Library of Congress.

[6] Chrispeels, J. H., Brown, J. H. & Castillo, S. (2000). School Leadership Teams: Factors that influence their development and effectiveness. *Understanding Schools as Intelligent Systems,* Vol. 4, 39-73, JAI Press.

[7] Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance*. Thousand Oaks, CA: Corwin Press. Retrieved on 2008-07-24 from http://findarticles.com/p/articles/mi_m0JSD/is_8_61/ai_n6191437.

[8] Feldman, J., & Tung, R. (2001). *Using data-based inquiry and decision making to improve instruction*. ERS Spectrum 19(3), 10-19.

[9] Wayman, J. C. , Stringfield, S. & Yakimowski, M. (2004). Software Enabling School Improvement Through Analysis of Student Data. *Report No. 67*. John Hopkins University. United States.

[10] Armstrong, J., & Anthes, K. (2001). *How data can help*. American School Board Journal 188(11), 38-41.

[11] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. SPSS White paper-technical report CRISPWP-0800, SPSS Inc.