4

Mining Enrolment Data Using Predictive and Descriptive Approaches

Fadzilah Siraj and Mansour Ali Abdoulha
Applied Sciences, College of Arts & Sciences, Universiti Utara Malaysia
Malaysia

1. Introduction

In recent years, the technology of database has become more advanced where huge amount of data is required to be stored in the databases, and the wealth of information hidden in those datasets has been realized by business people as a useful tool for making business strategic decisions. Data mining then attract more attention as it promises to extract valuable information from the raw data that businesses can use to increase their profitability through a profitable decision-making process.

Data mining is used to describe knowledge in databases; it is a process of extracting and identifying useful information and subsequent knowledge from databases using statistical, mathematical, artificial intelligence and machine learning technique (Efraim *et al*., 2007). Data mining applies modern statistical and computational technologies in its quest to expose useful pattern 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 data mining comes from its unique design by combining techniques from machine learning, pattern recognition, and statistics to automatically extract concepts, and to determine the interrelations and patterns of interest from large databases (Edelstein, 1997).

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 such as industries and government agencies. To remain competitiveness among educational field, these organizations need deep and enough knowledge for a better assessment, evaluation, planning, and decision-making. Majority of the required knowledge that has been 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.

Data mining 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. Data mining 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 (Luan, 2004).

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 data mining 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 data mining 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 approaches 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 namely, ***Neural Network, Logistic Regression*** and the ***Decision Tree***. The study shows that ***Neural Network*** obtains the highest results accuracy among the three techniques.

2. Data mining tasks

The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data (Chang & Hsu, 2005). The tasks of data mining can be modeled as either ***Predictive*** or ***Descriptive*** in nature (Dunham, 2003). A ***Predictive*** model makes a prediction about values of data using known results found from different data while the ***Descriptive*** model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. ***Predictive*** model data mining tasks include classification, prediction, regression and time series analysis. The ***Descriptive*** task encompases methods such as Clustering, Summarizations, Association Rules, and Sequence analysis (Fig. 1).

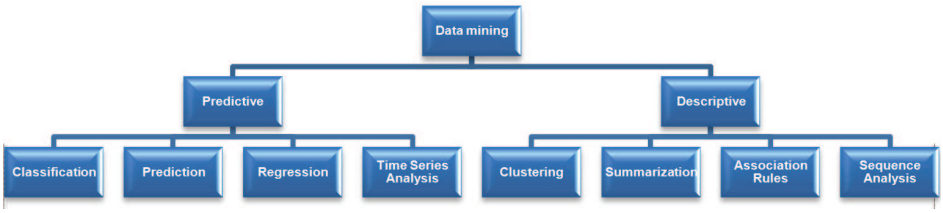


Fig. 1. Data mining tasks and models

Among ***Predictive*** models, Classification is probably the best understood of all data mining approaches. Three common characteristics of classification tasks are

- Learning is supervised
- The dependent variable is categorical
- The model built is able to assign new data to one of a set of well-defined classes.

For example, given classes of patients that corresponds to medical treatment responses; the form of treatment to which a new patient is most likely to respond to is identified (Stephens & Pablo, 2003). Unlike a classification model, the purpose of Prediction model is to determine the future outcome rather than current behaviour. Its output can be categorical

or numeric value. For example, given a prediction model of credit card transactions, the likelihood that a specific transaction is fraudulent can be predicted.
Another ***Predictive*** model known as statistical Regression is a supervised learnng technique that involves analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the development of a model that can predict these attribute values for new cases. For example, given a data set of credit card transactions, a model that can predict the likelihood of fraudulence for new transactions can be built. Prediction applications with one or more time-dependent attribures are called time-series problems. ***Time series analysis*** usually involves predicting numeric outcomes such as the future price of individual stock (Roiger & Geatz, 2003).
The second approach of data mining is known as ***Descriptive*** method. Descriptive data mining is normally used to generate frequency, cross tabulation and correlation. ***Descriptive*** method can be defined to discover interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data (Marco & Gianluca, 2005). In education, studies McNamarah (2005) used ***Descriptive*** to determine the demographic influence on particular factors. ***Summarization*** maps data into subsets with associated simple descriptions (Dunham, 2003). Basic statistics such as Mean, Standard Deviation, Variance, Mode and Median can be used as ***Summarization*** approach.
In ***Clustering***, a set of data items is partitioned into a set of classes such that items with similar characteristics are grouped together. ***Clustering*** is best used for finding groups of items that are similar. For example, given a data set of customers, subgroups of customers that have a similar buying behaviour can be identified.
Associations or *Link Analysis* are used to discover relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. i.e. to what extent one item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model. These relations may be associations between attributes within the same data item like ('Out of the shoppers who bought milk, 64% also purchased bread') or associations between different data items like ('Every time a certain stock drops 5%, it causes a resultant 13% in another stock between 2 and 6 weeks later'). ***Association Rules*** is a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored (Roiger & Geatz, 2003).
The investigation of relationships between items over a period of time is also often referred to as ***Sequence Analysis*** (Han & Kamber, 2001). ***Sequence Analysis*** is used to determine sequential patterns in data (Dunham, 2003). The patterns in the dataset are based on time sequence of actions, and they are similar to association data, however the relationship is based on time. In Market Basket analysis, the items are to be purchased at the same time, on the other hand, for ***Sequence Analysis*** the items are purchased over time in some order.

3. Data mining in education

It is highly necessary to determine that data mining techniques 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 ***Association Rules*** or ***Link Analysis*** which mostly used in telecommunication companies to understand groupings associated with starting points (Luan, 2001). Furthermore, prediction from Data Mining offers the university 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 activities on those students projected to experience such graduating difficulties.

The university's data can be used to suggest solutions to a wide range of educational challenges. Seifert (2004) indicated that Data Mining can be used to explore differences, explore growth over time, evaluate programs, and to identify the root causes of problems in education as one of the many ways data can be used. A study by Chrispeels, Brown and Castillo (2000) 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. Kennedy (2003) 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. Feldman and Tung (2001) observed that frequent usage of data in schools has metamorphosized into a more professional culture. Educators in their study have become greater collaborators during decision-making process, and school business consequently has become a less "privatized" one. Wayman, Stringfield and Yakimowski (2004) noted that school leaders who were involved in the use of data often developed a mindset of being responsible for their own destiny, increasingly able to find and use information to inform the school improvement. Armstrong and Anthes (2001) noticed that the use of data has helped in raising expectations of teachers on their 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 Data Mining concept, decision makers (management) in the educational sectors will definitely find their jobs easier.

Although computers supporting knowledge management have been widely used in fields such as business, Thorn (2001) observed that schools presented difficult technical problems due to the variety of data needs and usage at schools. School data is always found to be in different forms and places, making it more difficult to organize the databases effectivively. In addition, Thorn described a case study where a particu;ar district was ready to implement data that has to do with decision-making, however some technological barriers served as an hindrance to the process. Recent technological advances inevitably helps schools to overcome problems resulting from such technological barriers.

Data mining techniques, for example have been used to predict student performance in certain courses, to forecast the lecturer performance at the university and others. Indirectly these techniques contribute towards a better quality education management as well as assisting the education institution managing the administrative task effectively. Schools for example will soon have a variety of affordable, efficient computing tools to aid in the data mining process (Wayman & Stringfield, 2006).

4. Methodology

The CRISP-DM methodology suggested by Chapman *et al*. (2000) was utilized in this study. This methodology involves six phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment as shown in Fig. 2.

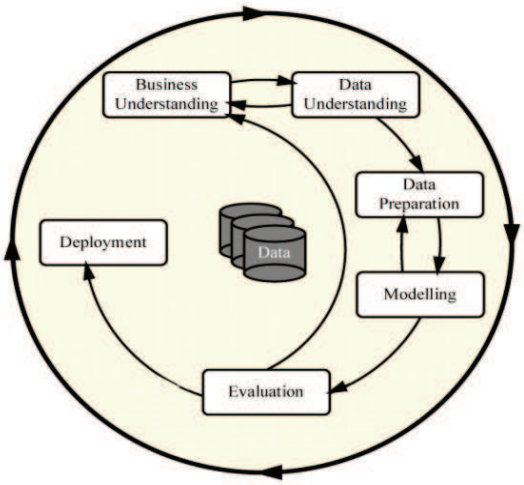


Fig. 2. Steps of CRISP-DM Methodology (Adopted from Chapman *et al.,* 2000).

In this study, as the possible areas of tests depends on the data available, and the detailed business objective cannot be identified until the data was 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.

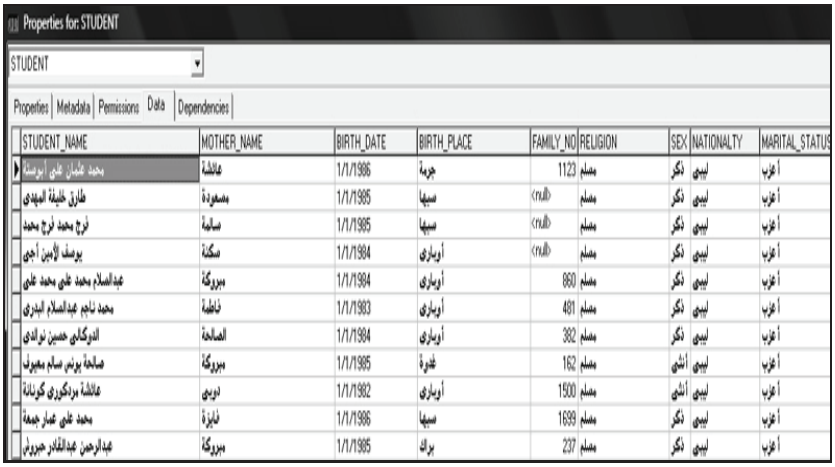
4.1 Business understanding

The first phase of CRISP-DM is business understanding which focuses on project objectives and requirement from a business perspective, and converting this knowledge into a data mining problem definition as well as designing a preliminary plan to achieve the objectives. To identify the research gap and the potential problems, literature study was carried out and relevant works have been identified. In this study, research related to educational data and factors affecting students' enrollment were sought and suitable mining algorithms for ***Predictive*** and ***Descriptive*** purposes were also selected.

4.2 Data understanding

The second phase is data understanding which begins with initial data collection. At this point, the data collected from the respondents needs to be checked and understood. In order to be familiar with the data, the next step in data understanding is to identify data quality problem, get some insights about the data and detect interesting subsets to form hypotheses so as to uncover the hidden information within the data collected for this study.

A total of 8510 students' enrollment from 1998 to 2006 was collected. An original student's main table includes 38 attributes with 8 numerical attributes and the others were of categorical type. Part of the original data is shown in Table 1.


Properties for: STUDENT

STUDENT

Properties | Metadata | Permissions | Data | Dependencies

STUDENT_NAME	MOTHER_NAME	BIRTH_DATE	BIRTH_PLACE	FAMILY_NO	RELIGION	SEX	NATIONALTY	MARITAL_STATUS
محمد عثمان علي أبوسنة	عائشة	1/1/1986	جرمة	1123	مسلم	ذكر	ليبي	أعزب
طارق خليفة المهدي	مسعودة	1/1/1985	سبها	<null>	مسلم	ذكر	ليبي	أعزب
فرج محمد فرج محمد	سالمة	1/1/1985	سبها	<null>	مسلم	ذكر	ليبي	أعزب
يوسف الأمين آجي	سكنة	1/1/1984	أوباري	<null>	مسلم	ذكر	ليبي	أعزب
عبدالسلام محمد علي محمد علي	مبروكة	1/1/1984	أوباري	860	مسلم	ذكر	ليبي	أعزب
محمد ناجم عبدالسلام البدري	فاطمة	1/1/1983	أوباري	481	مسلم	ذكر	ليبي	أعزب
الدوكالي حسين توالدي	الصالحة	1/1/1984	أوباري	382	مسلم	ذكر	ليبي	أعزب
صالحة يونس صالم مجبوب	مبروكة	1/1/1985	غدوة	162	مسلم	أنثى	ليبي	أعزب
عائشة مرادكوري كونانة	دوبي	1/1/1982	أوباري	1500	مسلم	أنثى	ليبي	أعزب
محمد علي عمار جمعة	نايزة	1/1/1986	سبها	1699	مسلم	ذكر	ليبي	أعزب
عبدالرحمن عبدالقادر هيروبي	مبروكة	1/1/1985	براك	237	مسلم	ذكر	ليبي	أعزب

Table 1. Sample of Student Data

As a result of preprocessing phase, the total number of data was reduced to 6830. In this phase, the data quality problem has been identified. The data were loaded into SAS version 9.13, checking for attributes to be analyzed and it was further processed in the next phase.

4.3 Data preparation

Data preparation concerns all activities needed to construct the final dataset for modeling purposes. The tasks are most likely to be carried out multiple times and may not be in any prescribed manner. Different datasets tend to expose new issues and challenges. With the goals in mind, it is important to choose the right data mining algorithms, techniques and tools which are expected to give best results with the provided data. Dependencies among different subsets of attributes are expected to be exhibited by different subsets of data. Most often, not all variables are used in analyzing and modelling process. This phase was conducted repetitively for determining suitable attributes to be used as predictors and target (output).

To get full insight of data distribution and in identifying outliers, descriptive analysis was conducted for exploratory purposes. In this study, the Cluster number was assumed as a target and other attributes such as demographics information, qualification upon entry and examination results were considered as predictor variables.

4.4 Modeling

During the modeling phase, modeling techniques were selected and applied to the dataset used in the study. This phase include selecting an appropriate modeling technique, building the models and followed by assessment of model. Subsequently, the model selection involves selecting appropriate techniques for the problem, refine the models whenever is necessary in order to meet the requirements and other constraints. After reviewing data

mining techniques, the ***Descriptive*** approaches employed in the study were identified, namely the ***Summarization*** and ***Clustering*** methods. Since the aim of the experiment was to study the patterns and getting some information within the enrollment data, no specific target has been identified. To this end, clusters were generated by ***Clustering*** method, and later used as target or output for ***Predictive*** approach.
Once the ***Descriptive*** methods has been identified, the next step was to identify the ***Predictive*** methods to be utilized in the next phase of empirical study. The identified approaches include ***Regression, Decision Trees and Neural Network***. In addition, comparison between these supervised approaches was also conducted to get some insight about the strength and weaknesses of each approach since one of the aims of the study was to determine whether these methods were well suited for extracting the required knowledge. As a result, the predictive method will be able to predict in which cluster does the future student will fall into based on the enrollment information.

4.4.1 Descriptive

Initially, ***Descriptive Statistics*** was carried out to investigate the nature of the dataset and the distribution of each attribute. Frequency tables were generated and the correlation analysis was also conducted to determine the relationship among the attributes, including Cross Tabulation Analysis (contingency tables). Cross Tabulation Analysis displays the relationship between two or more categorical (nominal or ordinal) variables.
Clustering Analysis was performed based on 4 clusters and the analysis conducted at this stage was based on the results obtained from Neural Connection software. For ***Clustering Analysis***, Kohonen network was used (Fig. 3) assuming the clusters, or classes, were formed from patterns that share common features and similar patterns have been grouped together.
Kohonen networks are usually one or two dimensional grids of artificial neurons, or nodes, where every node in the grid is connected to all the inputs. As the output comes directly from the grid of neurons (known as the Kohonen layer), Kohonen networks have no separate output layer. Each artificial neuron is linked to each input with a weight, and can be thought of as being at a point in the input data space.

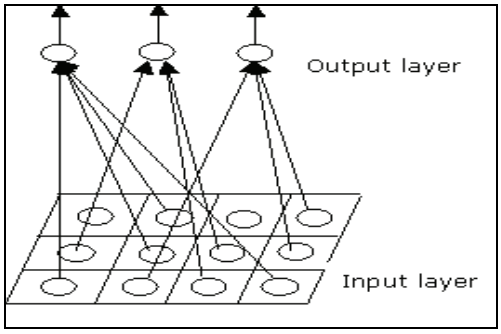


Fig. 3. Kohonen Network

Prior to training, these weights are set to initial values. Each output node at the output layer has an activation function and the output node with the best output wins the competition. The output node is identified as the output (or cluster) for that particular pattern. To enable

the Kohonen layer to group similar patterns, a neighbourhood of artificial neurons around the winning artificial neuron is also altered to be more like the input pattern. This is equivalent of moving the node towards the position of a pattern in the input data space. After a number of passes through the Kohonen layer, different areas of the Kohonen layer will respond to different types of patterns within the dataset (Everitt, 1993).

4.4.2 Predictive

As the clusters were generated through Kohonen networks, these clusters were then used as output for the predictive methods. Three predictive techniques, viz. ***Regression, Decision Tree,*** and ***Neural Network*** were employed to test the accuracies of the predictive models based on clusters.

The ***Regression*** analysis model is also known as one of the most useful tools in quantitative analysis phase of the decision-making process (Marquez *et al*, 1991). It is generally used to predict future values based on past values by fitting a set of points to a curve (Dunham, 2003). The simplest form of a regression model contains a dependent variable called outcome variable and single independent variable call factor. Logistic Regression is part of a category of statistical models called generalized linear model. This model includes ordinary regression and ANOVA as well as multivariate statistics such as ANOVA and loglinear regression. Logistic Regression allows one to predict a discrete outcome, such as group membership from set of variable that may be continuous discrete or a mix of any of this.

Decision Tree is a predictive model with tree or hierarchical structure, and commonly used in classification and prediction methods. It consists of nodes, which contained classification questions, and branches, or the results of the questions. At the lowest level of the tree - leave nodes - the label of each classification is identified. Typically, like other classification and prediction techniques, the ***Decision Tree*** begins with exploratory phase. Its algorithm will try to find the best-fit criteria to distinguish one class from another. The major concerns of this techiques are the quality of the classification problems as well as the appropriate number of levels of the tree. Some leaves and branches need to be removed in order to improve the performance of the decision tree. This step is also called tree pruning. The experiments using ***Decision Tree*** were conducted in parallel with the ***Regression Analysis*** and ***Neural Network*** modelling, and the algorithm used in ***Decision Tree*** is C4.5. Standard decision tree learners such as C4.5 increase the nodes in depth-first order (Quinlan, 1993) while in best-first decision tree the "best" node is expanded first. The "best" node is the node whose split leads to maximum reduction of impurity (e.g. Gini index) among all nodes available for splitting (Shi, 2006).

Neural Network model known as multi layer perceptron with back propagation algorithm was used to establish a prediction model (Fig. 4). The distribution of data for training, validation and test set were evaluated to determine the suitable composition for obtaining a good prediction model. Some back propagation parameters were also investigated to obtain a suitable Neural Network prediction model (Sirikulvadhana, 2002). In general, there are three types of activation functions that are commonly used in neural network, namely Threshold function, Piecewise-linear function and Sigmoid function. Different from other learning algorithms, backpropagation algorithm works, or learns and adjusts the weights backward, which simply mean that it predicts the weighted algorithms by the input from the output.

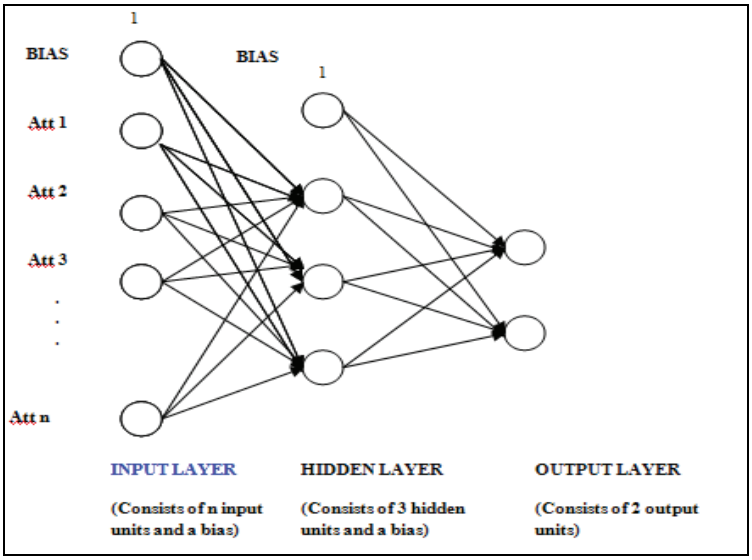


Fig. 4. The architecture of Multilayer Perceptron

4.5 Evaluation

In this phase, models were evaluated to assess the degree to which the model meets the business objectives and quality requirements. The steps involved include evaluating the results, reviewing the processes and determining the next steps. The evaluation for ***Regression, Neural Networks and Decision Tree*** was based on the classification accuracy, confusion matrix table, and classification table respectively.

Deployment

The last phase in CRISP-DM was the deployment. The knowledge from the model acquired from the experimental study were to be transferred for implementaion purposes.

5. 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, the distribution of students enrollment is shown in Fig. 5. Programs such as Dentistry (Sebha), Sport (Ghat), Arts (Tragen) and Sciences (Tragen) have small enrollment figures, ranging from 1% to 4% with number of students less than 350 over 8510 of all university students' population. This indicates that the university should put this fact in consideration in coming years as to which location has low students' population. However, if the faculties are grouped by the cities, Sebha faculties (Sciences, Arts, Medicine, Dentistry, Law and Agronomy) contributed to 55% of university population.

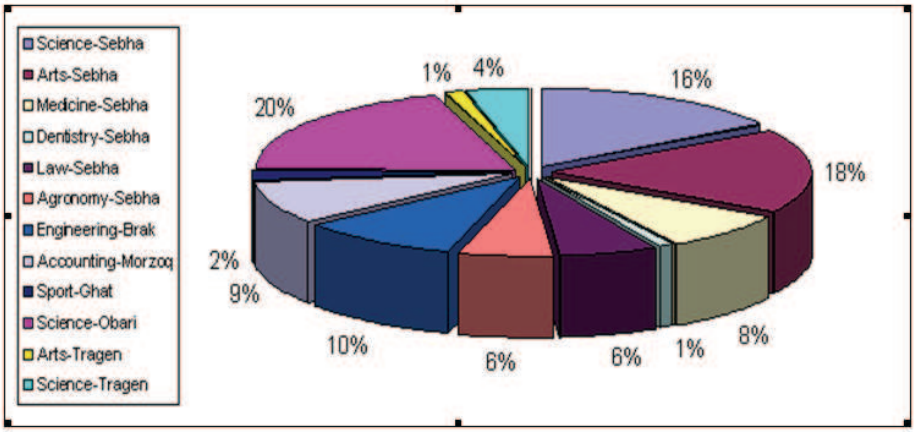


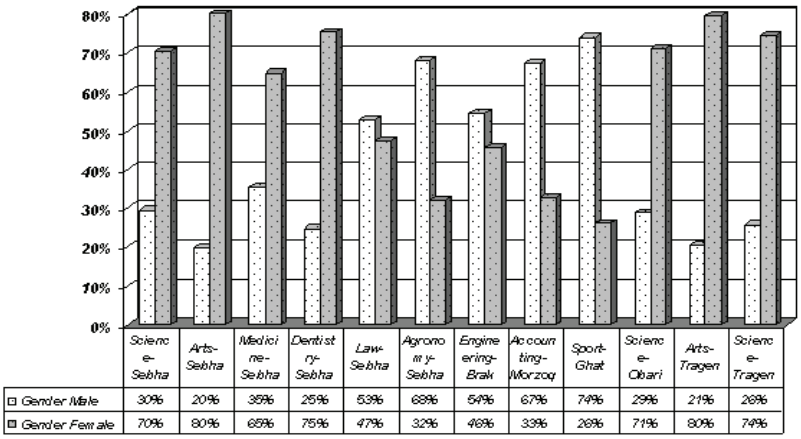
Fig. 5. Distribution of Student Population

The ***Descriptive*** statistics, particularly ***Cross Tabulation Analysis*** was carried out to discover the relationship between the attributes (Fig. 6). Based on the results shown in Fig. 5, majority of the registered students were female with ratio of 58% in all university population and almost 42% were male. Program such as Sciences (Sebha, Obari and Tragen), Arts (Sebha and Tragen), Medicine (Sebha) and Dentistry (Sebha) were 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. 7.

Based on results exhibited in Fig. 8 and 9, 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. 8). In addition, Agronomy (Sebha) program was not preferred by female students. Fig. 9 indicates students that have been expelled from continuing Engineering (Brak) and Sciences (Sebha) programs Arts (Sebha) degree program.

The results shown in Fig. 10 indicate that more female (63%) enrolled at the university compared to male (37%). In other words, the male enrollment rate was almost half of the female rate. Most of female students (69% to 81%) undertook Arts (Sebha and Tragen), Science (Tragen, Obari and Sebha), Dentistry (Sebha), and Medicine (Sebha). The least number of female students undertook Sport (Ghat), Accounting (Morzoq) and Agronomy Sebha. On the other hand, majority of the male students were enrolled in the programmes such as Sport (Ghat), Accounting (Morzoq) and Agronomy (Sebha).



	Science-Sebha	Arts-Sebha	Medicine-Sebha	Dentistry-Sebha	Law-Sebha	Agronomy-Sebha	Engineering-Brak	Accounting-Morzoq	Sport-Ghat	Science-Obari	Arts-Tragen	Science-Tragen
Gender Male	30%	20%	35%	25%	53%	68%	54%	67%	74%	29%	21%	26%
Gender Female	70%	80%	65%	75%	47%	32%	46%	33%	26%	71%	80%	74%

Fig. 6. Faculty with Respect to Gender

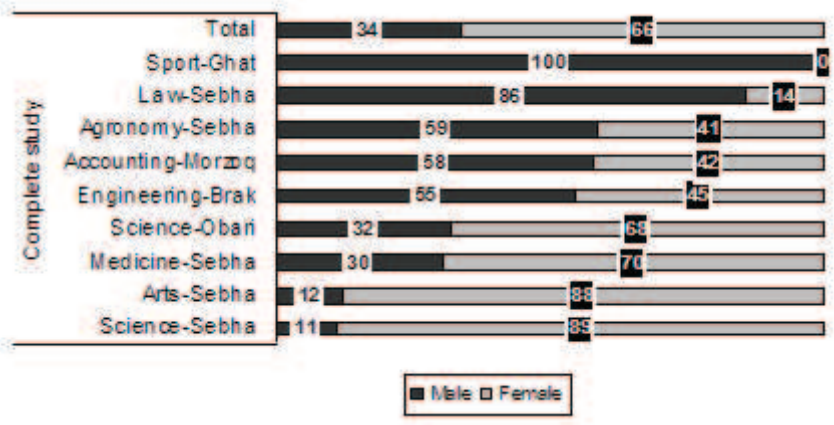


Fig. 7. Faculty with Respect to Student Status (*Complete Study)* and Gender

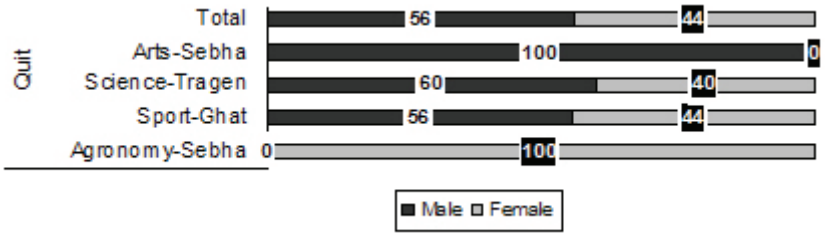


Fig. 8. Faculty with Respect to Student Status (*Quit*) and Gender

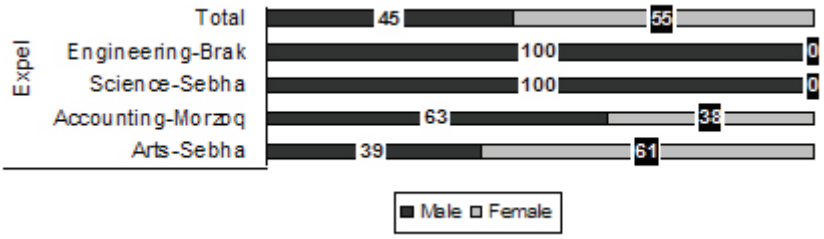


Fig. 9. Faculty with Respect to Student Status (*Expel*) and Gender

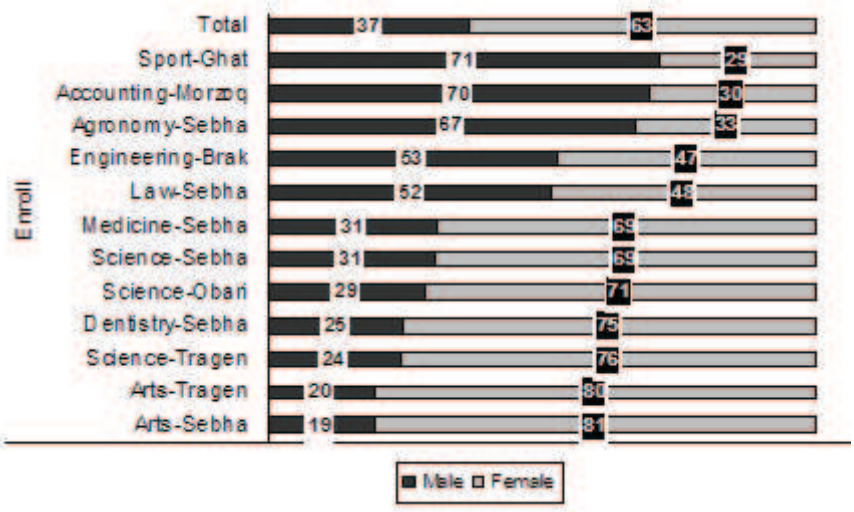


Fig. 10. Faculty with Respect to Student Status (*Enroll*) and Gender

Having performed ***Cross Tabulation Analysis***, the clustering network using Kohonen network has been carried out. As a result, 3 clusters have been identified (***Cluster 0***, ***1*** and **2**). Further investigation was performed to carry out in order 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).

	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).			

Table 2. The correlation between enrollment attributes and clusters

The main aim of clustering is to group cases based on its similarities. In addition, each ***Cluster*** has its own characteristic, which can be analyzed based on faculties using statistical approach. In order to determine the meaning of each ***Cluster***, cross tabulation analysis was carried out, and the ***Clusters*** analyzed based on faculties are shown in Fig. 11. Results in Fig. 11 indicates that more male in ***Cluster 0*** and **2**. On the other hand, female students has the tendency to fall into ***Cluster 1***.

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 relationship between the clusters based on Faculty and Gender are shown in Table 3. As for the faculty with respect to gender and cluster, clearly ***Cluster 1*** comprises of female students undertaking Arts, Sciences, Dentistry and Medicine. Male students who undertook Sports were most likely to fall into ***Cluster 3*** rather than 0. ***Cluster 3*** is more inclined to male students who undertook Law, Sport, and Sciences(Tragen). ***Cluster 0*** does not show any clear pattern.

When further analysis was performed on the clusters, it is intersting to note that the cluster is able to distinguish between Libyan and non-Libyan students. In addition, some rules with regard to faculty and nationality can also be extracted.

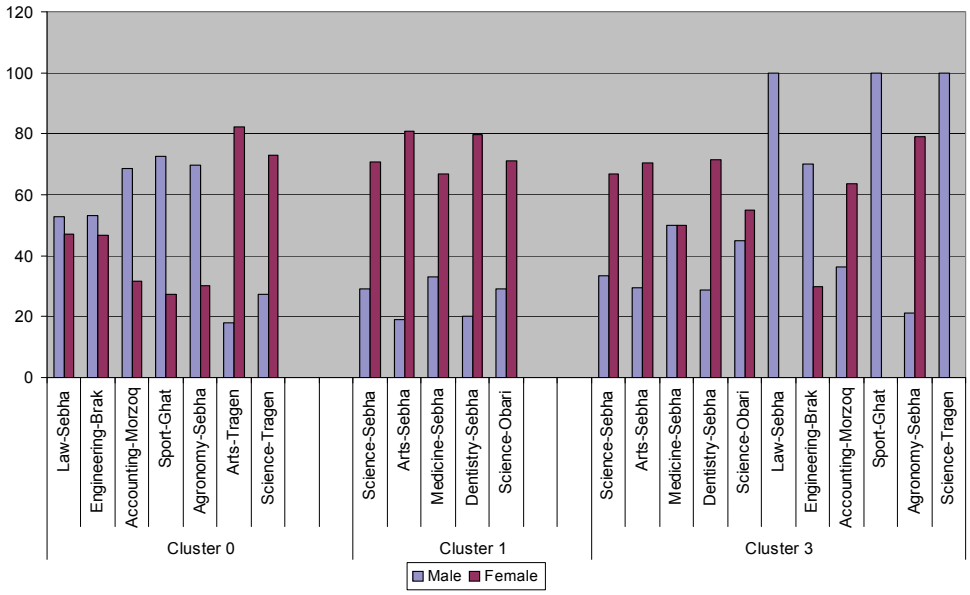


Fig. 11. Faculties with Respect to Gender for Each Cluster

Cluster 1		
Law-Sebha	Male	Almost the same
	Female	
Engineering-Brak	Male	Almost the same
	Female	
Accounting-Morzoq	Male	More
Sport-Ghat	Male	More
Agronomy-Sebha	Male	More
Arts-Tragen	Female	More
Science-Tragen	Female	More

Cluster 2		
Science-Sebha	Female	More
Arts-Sebha	Female	More
Medicine-Sebha	Female	
Dentistry-Sebha	Female	More
Science-Obari	Female	More

Cluster 3		
Science-Sebha	Female	More
Arts-Sebha	Female	More
Medicine-Sebha	Male	The Same
	Female	
Dentistry-Sebha	Female	More
Science-Obari	Male	Almost the same
	Female	
Law-Sebha	Male	All
Engineering-Brak	Male	More
Accounting-Morzoq	Female	More
Sport-Ghat	Male	All
Agronomy-Sebha	Female	More
Science-Tragen	Male	All

Table 3. The Relationship between the Cluster with Respect to Faculty and Gender

Cluster 0		
Law-Sebha	Libyan	All
Engineering-Brak	Libyan	Almost All
Accounting-Morzoq	Libyan	All
Sport-Ghat	Libyan	Almost All
Agronomy-Sebha	Libyan	Almost All
Arts-Tragen	Libyan	All
Science-Tragen	Libyan	Almost All

Cluster 1		
Science-Sebha	Libyan	Almost All
Arts-Sebha	Libyan	Almost All
Medicine-Sebha	Libyan	Almost All
Dentistry-Sebha	Libyan	All
Science-Obari	Libyan	All

Cluster 2		
Science-Sebha	Lebanese	More
Arts-Sebha	Palestinian	Almost the same
	Sudanese	
Medicine-Sebha	Palestinian	More
Dentistry-Sebha	Palestinian	Almost the same
	Iraqi	
Science-Obari	Palestinian	More
Law-Sebha	Syrian	Almost the same
	Chadian	
Engineering-Brak	Sudanese	More
Accounting-Morzoq	Sudanese	Almost all
Sport-Ghat	Tunisian	All
Agronomy-Sebha	Sudanese	Almost all
Science-Tragen	Sudanese	All

Table 4. The Relationship between the Cluster with Respect to Faculty and Citizenship

If the student is Libyan, and undertaking Law, he/she falls under ***Cluster 0***. This is also true for Accounting-Morzoq, Sport-Ghat, Agronomy-Sebha, Art-Tragen, and Science-Tragen. If the student is Libyan and taking Arts at Sebha, he/she falls in ***Cluster 2.*** This is also true for Libyan students at Sebha who undertook Medicine and Dentistry. However, those undertook Sciences program are from Obari and Sebha. Other international students fall into ***Cluster 3***. The relationship between the clusters based on faculty and citizenship are shown in Table 4. It is very obvious that Libyan students mostly fall into ***Cluster 0*** or ***1***.
The overall result of determining the characteristics of each ***Cluster*** and comparison between all clusters is shown in Table 5. The results exhibited in Table 3 indicate that some faculties are common to all clusters (for example Faculty of Sciences and Arts) whereas some have unique characteristics. For example, if the students undertake Arts at Tragen, the students fall into ***Cluster 0***, otherwise they fall into etiher ***Cluster 1*** or ***Cluster 2***. Students who undertook Medicine and Dentistry fall either into ***Cluster 1*** or ***Cluster 2***. On the other hand, students who undertook degree programs such as Law, Engineering, Accounting, Sport and Agronomy falls into ***Cluster 0*** or ***Cluster 2***. The proportion of male to female students in ***Cluster 0*** is nearly the same (52% and 48%), most of them stayed in University's residence and also 90% of them were admitted through government's process. 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 undertook 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 Non-residence students were selected through university selection process.

Variables	*Cluster 0*		*Cluster 1*		*Cluster 2*	
FACULTY	**Degree**	**Place**	**Degree**	**Place**	**Degree**	**Place**
	Sciences	Tragen	Sciences	Sebha	Sciences	Sebha
			Sciences	Obari	Sciences	Obari
					Sciences	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%

Table 5. Clusters characteristic with respect to predictors' variables

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%). After the ***Decision Tree*** analysis was performed, the accuracy for training and validation is high, for training is 99.77% and validation is also 99.77%. Like ***Regression Analysis*** results, ***Faculty*** and ***Nationality*** are two important variables in ***Decision Tree*** 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. 12 illustrates the lift chart for the three prediction models based on clustering results as the target.

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. The empirical results and the analysis indicate that the ***Descriptive*** and ***Predictive*** methods based on clusters have revealed some characteristics and also uncover more hidden information within Sebha University enrollment data.

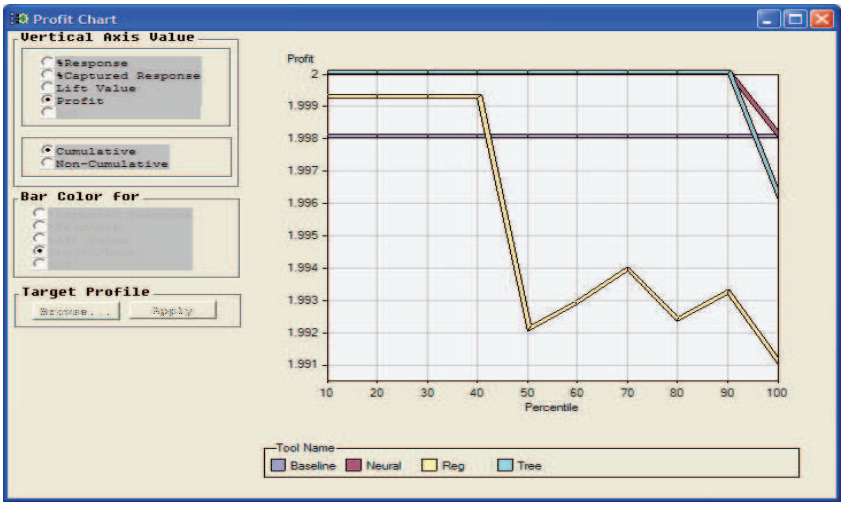


Fig. 12. Comparison of Accuracies Between Regression, Neural Network and Decision Tree Predicton Techniques

6. 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, it is around 15 kilometre. Descriptive statistics and correlation analysis defined two attributes 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) were 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 degrees in future.

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.

7. References

Armstrong, J. & Anthes, K. (2001). How data can help, *American School Board Journal*, Vol. 188, No. 11, pp. 38-41.

Brown, J. D. (2007) Neural Network Prediction of Math and Reading Proficiency as Reported in the Educational Longitudinal Study 2002 Based on Non- Curricular Variables, Ph.D Dissertation, Duquesne University.

Chang, H. C. & Hsu, C.C. (2005). Using Topic Keyword Clusters for automatic Document Clustering. *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, Kota Kinabalu, Sabah.

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.

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, pp. 39-73, JAI Press.

Dunham, M. H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Earl, L. & Katz, S. (2002). Leading School in a data-rich world, In: *Second International Handbook of Leadership and Administration*, Leithwood, K & Hallinger, P., pp. 653-696, Kluwer, Dordrecht.

Edelstein, H. (1997). Data mining: Exploring the hidden trends in your data. *DB2 Online Magazine*. Available: http:.//www.db2mag.com (URL)

Efraim, T.; Jay, E. A.; Tin-Peng, L. & Ramesh, S. (2007). *Decision Support and Business Intelligent Systems*, Pearson Education.

Everitt, B. S. (1993). Cluster analysis (2nd Ed.), Edward Arnold, London.

Feldman, J. & Tung, R. (2001). Using data-based inquiry and decision making to improve instruction, *ERS Spectrum*, Vol. 19, No. 3, pp. 10-19.

Golding, P. & McNamarah, S. (2005). Predicting academic performance in the school of computing & information technology (SCIT). *35th ASEE/IEEE Frontiers in Education Conference.* Indianapolis.

Han, J. & Kamber, M. (2001). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems), Academic Press, San Diego

Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance*. Thousand Oaks, CA: Corwin Press. Available at: http://findarticles.com/p/articles/mi_m0JSD/is_8_61/ai_n6191437.

Luan, J. (2001). Data Mining as Driven by Knowledge Management in Higher Education-Persistence Clustering and Prediction, presented at 2001 SPSS Public Conference, UCSF.

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

Marco, R. & Gianluca, C. (2005). Data Mining Applied to Validation of Agent Based Models, *Proceedings of Ninteenth European Conference on Modelling and Simulation*, RIFA.

Marquez, L.; Hill, T.; Worthley, R. & Remus, W. (1991). Neural network models as an alternative to regression, Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences, Vol. iv, pp. 129 – 135.

Ogor, E. N. (2007). Student academic performance monitoring and evaluation using data mining techniques. *Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007),* pp. 354-359.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publisher, New York.

Roiger, R. J. & Geatz, M. W. (2003). Data Mining: A Tutorial-Based Primer, Addison-Wesley, ISBN 0-201-74128-8, Boston.

Seifert, J. W. (2004). *Data mining: An overview*. Proceeding of CRS Report for Congress, Library of Congress, pp. 1-16.

Shi, J. (2006). Best-first Decision Tree Learning, *MSc. Thesis*, University of Waikato, New Zealand.

Sirikulvadhana, S. (2002). Data mining as a financial auditing tool. *MSc. Thesis in Accounting*, The Swedish School of Economics and Business Administration, retrieved on September, 12, 2008 from www.pafis.shh.fi/graduates/supsir01.pdf.

Stephens, S. & Pablo, T. (2003). Supervised and unsupervised data mining techniques for the life sciences. *Technical Report*, Oracle and Whitehead Institute, MIT, USA.Thorn, C. A. (2001). Knowledge Management for Educational Information Systems: What is the State in the Field?, *Eucation Policy Analysis Archieves*, Vol 9, Issu 47, retrieved July 22, 2008, from http://epaa.asu.edu/epaa/v9n47/.

Wayman, J. C.; Stringfield, S. (2006). Technology-Supported Involvement of Entire Faculties in Examination of Student Data for Instructional Improvement, *American Journal of Education*, Vol. 112, pp. 1-23.

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