Gezira j. of Eng. & applied. Sci. 12-(1): 44-58 (2017)

Use Data Mining for Evaluation University Student Results

Sally Dafa-Allah AwadElkarim Aboalgasim [1], Mujahed Ahmed Ali Ahmed [2]

[1]Department Computer Engineering, Faculty of Engineering &Technology, University of Gezira, Wed Medani, Sudan.
[2]National Cancer Institute, Gezira State, Wed Medani, Sudan.

ABSTRACT

The student academic performance is based on different factors such as social and technical personality. The aim of this paper is to analyze students' results in different years using data mining in order to extract important information, then using this information in decision making process. This paper used classification of supervised learning algorithm with its technologies in the area of decision tree to identify the primary influence. The implementation showed the largest influence on students' results according to factors like batches, major, admission type and sex. It also showed that the results of batches admission in year 2007 and 2011 had the largest number of dismissed students with a percentage of 13% and 12% respectively.

INTRODUCTION

The amount of data continues to grow at huge rate even though the data storage are already vast. The main challenge is how to make the database competitive business advantage by converting seemingly meaningless data into useful information. How this challenge is met is critical because universities are increasingly relying on effective analysis of information simply to remain competitive. A combination of new techniques and technology is emerging to help sort through the data and find useful competitive data (Ian H. Witten 2005).

By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in a database and investigated from different angles. Large databases thereby serve as rich reliable sources for knowledge generation and verification. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database systems and machine learning. Companies in many industries also take knowledge discovery as an important area with an opportunity of major revenue. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications (Fayyad, Piatetsky-Shapiro et al. 1996).

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is a process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (S.Linoff 2004). The Statistical Analysis System (SAS) Institute(Sato 2000) defines data mining as the process of selecting, exploring and modeling a large amount of data to uncover previously unknown data patterns for business advantages. Data mining refers to as knowledge discovery in databases. It means a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases (Fayyad, Piatesky-Shapiro et al. 1991) from business perspective, various data mining techniques are used to better understand user behavior, to meliorate the service provided, and to increase the business opportunities (Chen, Han et al. 1996).

There are increasing research interests in using data mining in education. This new emerging field is concerned with developing methods that extract knowledge from educational data. The data can be collected form historical and operational data that reside in the databases of educational institutes.

The student data can either be personal or academic. Data can also be collected from e-learning systems which have a large amount of information offered by most institutes. Many technologies can be used to support the implementation of knowledge management systems (KMS). These technologies range from simple to complex ones and could be classified into two types information technologies and web-based technologies (Mohammed M. Abu Tair February 2012).

Information technologies include the databases, decision support systems, intelligent agents and data mining. Whereas, web-based technologies consist of internet, intranet, email, groupware and many more. Each of these two technologies has different function and ability to support KM. However, studies on how these technologies can realize the KM processes in the context of institutes of higher learning (IHL) and how can they be integrated are at scarce. Indeed the successful development of KMS depends on the right selection of KM technologies. Data mining technology has been suggested as a promising technology for KMS implementation in IHLs despite the increasing number of other sophisticated technologies. This is due to the ability of the technology to extract useful knowledge from large volumes of data from databases which are part of KM initiatives. Moreover, nkydata mining technology is able to classify and predict future outcome based on previous data that could make an important contribution for KM processes; particularly in decision making process. In this case, applied data mining technology to huge amount of data could give greatest advantages to IHLs especially when IHLs are no longer simply providing knowledge to students, but are also able to manage and expand their existing knowledge for future reference. Even though data mining technology has been explored and evaluated IHLs in various countries, there is only a handful of research on how to capture knowledge from a source and manage the knowledge produced. In fact, studies on data mining have emphasized more on testing the ability of the algorithm to produce an accurate model, rather than on developing a comprehensive application system for education users to make decisions. In other words, there is a lack of studies on the integration of data mining technology in the context of KMS for IHLs (Muslihah Wook 2013).

Data mining helps to extract the knowledge from available dataset and should be created as knowledge intelligence for the benefit of the institution. Higher education does categorize the students by their academic performance. Many factors influence the academic performance of the student. The model is mainly focused on exploring various indicators that have an effect on the academic performance of the students.

The extracted information that describes student performance can be stored as intelligent knowledge for decision making to improve the quality of education in institutions. The knowledge stored is used for predicting the student's performance in advance (R. Sumitha and E.S. Vinothkumar, 2016).

The higher education institutions has potential knowledge such as academic performance of students, administrative accounts, potential knowledge of the faculty, demographic details of the students and many other information in a hidden form. The technique behind the extraction of the hidden knowledge is Knowledge Discovery process. Recently Data mining is widely used on educational dataset. Educational Data mining (EDM) has become a very useful research area (Crist´obal Romero,2010).

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal (Hem Jyotsana Parashar 2012).

Decision tree learning is used in statistics, data mining and machine learning. it uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. in decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making (Tom Mitchell,1997).

In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. The main idea of how to choose an attribute and the parameter of the basic structure of creating a decision tree is the same for most decision tree algorithms. The difference lies in how to select the attributes for the tree. Here the focus will be on the ID3 algorithm developed by Ross Quinlan in 1975.

METHODOLOGY:

In this application SQL server was used for database design and visual basic was used to formats (Dev Express Dev Extreme Demos 13.1). Figure (1) illustrate the data mining steps that used as a methodology for this paper. The classification model and the decision tree technique were used to develop an application for the University of Gezira, Faculty of Mathematical and Computer Sciences (FMCS).

The results of students who admission in year 2007 to 2011were used as dataset for designed program ,then the classification rules were built.

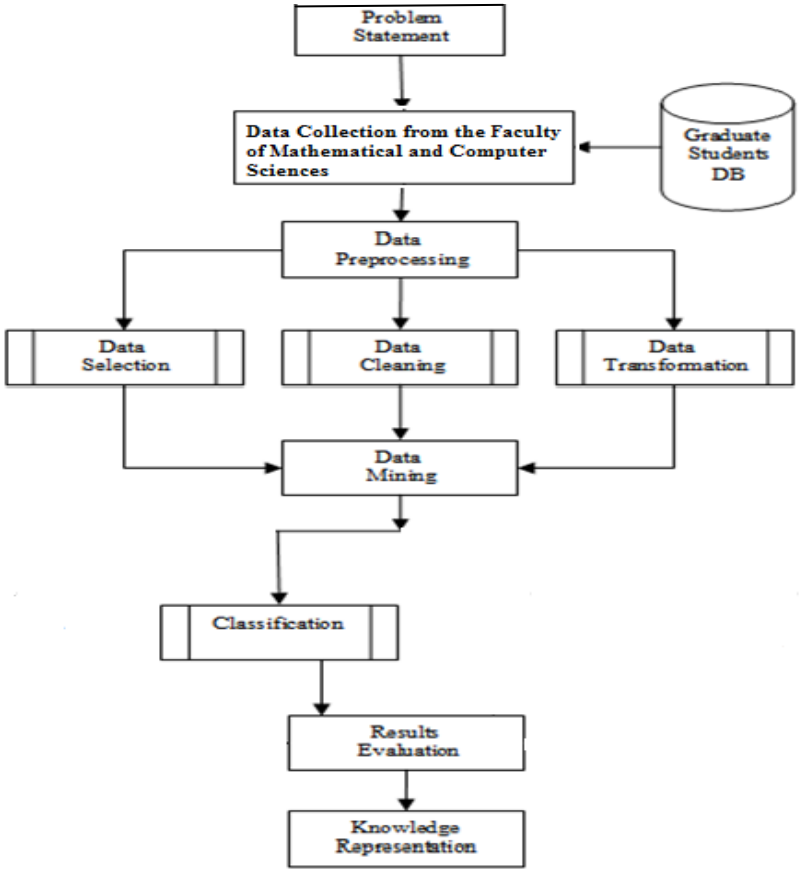

Figure (1) Data Mining steps and Methodology for Student Results

Classification rules discovered by the program may also give ideas to the university officials, in providing appropriate courses for students. After collecting data of students and their results from examination and registration office in FMCS they will be put in classification rules then properties transaction will be set for program design will be developed.

1. International Direct Distance Dialing (ID3):

The application in this paper used the classification of supervised learning, then the decision trees and ID3 algorithm were used to design and support data mining techniques. Decision trees technique is commonly used in data mining because their construction is cheap, their interpretation is easy, and their integration with database systems is also easy and for their good reliability.The basic steps behind ID3 algorithm are as follows:

1. Choose the best attribute(s) to split the remaining instances and make that attribute a decision node.
2. Continue doing this process to each node or sub node, and stop when:-
    a) All the instances have the same target attribute value.
    b) There are no more attributes.
    c) There are no more instances.
3. To determine the best attribute, look at the ID3 heuristic, ID3 splits attributes based on their entropy, and entropy is the measure of disinformation.
4. Entropy is minimized when all values of the target attribute are the same. For example the time will be short, then entropy = 0, entropy is maximized when there is an equal values choice for all target attributes (i.e. the result is random) (Andrew Colin, Dr. Dobbs 1996).

1.1 Calculation of entropy

$$\text{Entropy}(S) = \sum_{i=1}^{n} |Si|/|S| * \log2(|Si|/|S|) \quad (1)$$

Where:-

S = set of examples.

S_i = subset of S with value under the target attribute.

n = size of the range of the target attribute. (Tom Mitchell,1997)

ID3 algorithm will split attributes according to the lowest entropy, the entropy for all values of an attribute, are calculated as the weighted sum of subset entropies as follows:

$$\text{Entropy}(S) = \begin{cases} \sum_{i=1}^{n} -p_+ \log_2 p_+ & \text{if} p_i \geq 3 \\ \sum_{j=1}^{n} -p_- \log_2 p_- & \text{if} p_j < 3 \end{cases} \quad (2)$$

Where:

(p_+ ratio of positive examples over all examples). (CGPA≥3)

(p_- ratio of negative examples over all examples). (CGPA<3)

If there are more than two classes :

$$\text{Entropy}(s) = \sum_{i=1}^{n} - p(i) \log_2 (pi) \quad (3)$$

Where: n the number of classes.

Second information gain will be measured (which is inversely proportional to entropy) as follows:

$$\text{Entropy(S) - } = \sum_{i=1}^{n} |Si|/|S| \text{ Entropy(Si)} \quad (4)$$

1.2 The Data Set used in the Application

The data set used in the application contains all the student in the FMCS who graduated during the period between Dc, 2007 and Dc, 2011. The collected data are made separate in different eight tables, and these tables are presented as followed:

1. Students Table: shows the personal data collected for each student.
2. Student-Subject Table: describes the subjects taken by students in each semester.
3. Grade-Points table: deals with grades of each student in FMCS. Each semester has points calculated according to the grades obtained for each subject semester then the total points makes up GPA then CGPA will calculated by dividing GPA by the number of semesters completed.
4. Subjects Table: describes the subject which must belong to a department, with the numbers of credit hours, which are (2, 3 or 4).
5. Department Table: describes the type of department. There are two departments in FMCS the department there are: (statistics/computer science, Mathematical/computer science).
6. Jm Batches Table: describes the batches studied at FMCS and each field has a special number according of year of admission. These number are: (26 for student admission in year 2007, 27 for student admission in year 2008, 28 for student admission in year 2009, 29 for student admission in year 2010 and 30 for student admission in year 2011).
7. Teachers-Subject Table: shows teacher data. Each teacher must have a subject characteristic of teachers-subject.
8. Faculties Table: describes with faculties data each faculty must have a number to specific of faculties.

All these tables describe the data set which is used on implementation of the application. The relationships between these tables are described in Figure (2).

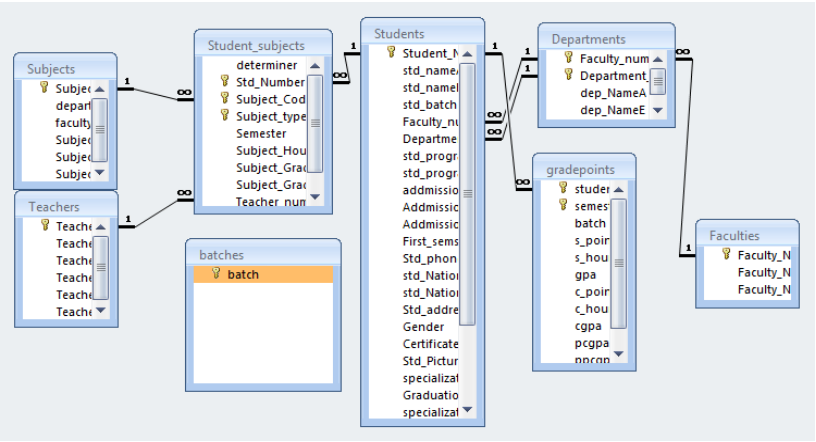

Figure (2): Relationship between Dataset Tables

1.3 Design of the Application

A supervised learning method is used to learn a rule or concept from a given set of positive and negative training data. And it was very popular data mining technique most commonly used in classification.

1.3.1 Top-Down Induction of Decision Trees (TDIDT)

Basic algorithm for TDIDT: (based on ID3)

Start with full dataset, then find the test that partition examples as good as possible.

For example item from same class, or otherwise similar, are put together. For each outcome of test, create child node, move examples to nodes according to outcome of test, repeat procedure for each child that is not "pure". Entropy for the training data example students graduated in 2011(batch (30)), E(S), is represented as E([49+,91-]) because of the 140 training examples 49 of them are **YES** and 91 of them are **NO**.

Entropy(S) = $-p_+ \log_2 p_+ - p_- \log_2 p_-$

E(S) = E ([49+, 91-]) = $(-49/140 \log_2 49/140) + (-91/140 \log_2 91/140)$ = **0.9341**

Calculate the information gain G(S, A) for each attribute A, where A is taken from the set (Admission-Type, Specialization, Gender) to determine root node, as shown in Table (1).

G(S, Admission-Type) = E(S) – [108/140 x E (Admission-Type =General) + 14/140 x E (Admission-Type = state) +15/140 x E (Admission-Type=special) +3/140 x E (Admission-Type=S-workers) + 0/140 x E (Admission-Type=award)]

= 0.9341–[108/140 x E (45+, 63-)+14/140 x E ([1+, 13-]) + 15/140 x E ([2+, 13-]) + 3/140 x E ([1+, 2-]) + 0] = **0.0804**

G(S, Specialization) = E(S) – [131/140 x E (Specialization=statistics) + 9/140 x E (Specialization = mathematical)]

= 0.9341 – [131/140 x E (44+, 87-) + 9/140 x E ([5+, 4-])] =**0.0088**

G(S, Gender) = E(S) – [57/140 x E (Gender=male) + 83/140 x E (Gender= female)]

= 0.9341 – [57/140 x E (17+, 40-) + 83/140 x E ([32+, 51-])] =**0.006**

Table (1): The Gain for Admission-Type, Specialization and Gender

Attribute	**Gain**
Admission-Type	**0.0804**
Specialization	0.0088
Gender	0.006

The root node was selected from the attributes with the biggest information gain, that was admission-type, Figure (3) showed the sub-attribute for the root node:

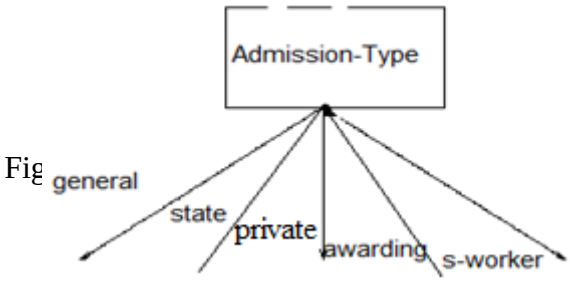

Figure (3): Sub-Attribute for Admission-Type

Then calculate the information gain G(S, A) for each attribute A, where A is taken from the set (general, state, special, s-worker, award) to determine sub node, that shown in Table (2) , where the selected node is gender for general and special admission type and the selected node is specialization for state and s-worker admission type.

Table (2): Information Gain of Admission Type

Attribute	**Gain**	
	Gender	specialization
general	**0.0047**	0.0001
state	0.0476	**0.3711**
special	**0.0851**	0
Awarding	0	0
s- worker	0.2516	**0.9183**

The greater information gain shown in order: admission type, the root of tree, and that it greater than Specification and gender. To illustrate and explain the results obtained from the application, a decision tree was created based on the ID3 algorithm as shown in Figure(4) and illustrate the complete decision tree for all attribute.

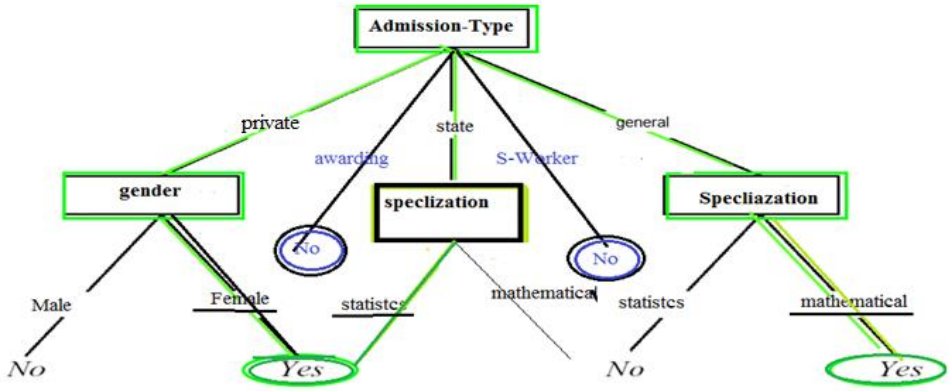

Figure (4): The complete Decision Tree after ID3 Algorithm

RESULTS

The application executes the classification techniques and using ID3 algorithm in the fields of Admission-Type, major and Gender.
The algorithm is applied to observe the result of student's performance in the batches which graduated from 2007 to 2011. And then classify the batches to admission-Type, specialization and gender and show the greater information gain to put the head of tree (the bigger parameter effect)
Data for students were graduated in 2011 was used as input to designed program. This algorithm uses the information gain to select the attributes and to build the tree. Figure (3) shows that the first two levels of the tree. In addition, each node of the tree can be used to extract a rule to predict students' final grades based on their activities. For example admission-type (general, state, worker, special, award), gender(male, female) and specification (Statistic, Mathematic). These parameters are used in algorithm.
The students were classified and separated into two groups: the student with CGPA less than 3.00 and student with CGPA equals or more than 3.00. The measurement from CGPA is 0.9341. The measurement according the gender is in Table (3).

Table (3): The Measurement According the Gender

Gender	Measurement
Male	0.8791
Female	0.9618

Then the data will be classify according to specification (Statistic, Mathematic) and CGPA, as shown in Table (4).

Table (4): The Measurement According the Specification and GPA

Specification	CGPA
Statistic	0.9208
Mathematic	0.9911

Table (5) illustrate the classification according to specification (Statistic, Mathematic), admission-type (general, state, worker, special, award).

Table (5): Classification According to Specification and Admission-type

	Measurement
CGPA and general	0.9799
CGPA and state	0.3711
CGPA and worker	0.9183
CGPA and special	0.5664
CGPA and award	0.0

Table (6) illustrate the classification according to admission-type (general, state, worker, special, award) and gender.

Table (6): Classification According to Admission-type and Gender

Admission type	Male	Female
general	0.9475	0.9921
state	0.5032	0.0
s-worker	1.0	0.0
special	0.0	0.7219
award	0.0	0.0

Table (7) illustrate the classification according to specification (Statistic, Mathematic), admission-type (general, state, worker, special, award).

Table (7): Classification According to Specification and Admission-type

Admission type	Specification statistic	Specification math
general	0.9794	0.9852
state	0.0	0.0
worker	0.0	0.0
special	0.5665	0.0
award	0.0	0.0

Table (8) illustrate the classification according to specification (Statistic, Mathematic and gender.

Table (8): Classification According to Specification and Gender

Specification	Male	Female
Statistic	0.819	0.9632
Mathematic	0.9183	0.09183

These rules may help the teachers to identify the most important activities to focus in order to improve their teaching style. The rules can also be employed by training managers and executives to provide with helpful information in resource planning and decision making.

DISCUSS.ION

Decision tree techniques and ID3 algorithms were used in this application to choose the best parameter among all parameters which passed to the algorithms. These parameters were admission type, major and gender. Then the batch was selected to calculate the percentage of this parameter to choose the effective parameter or the best parameter depending on the information gain. The best parameter will be taken and matched with other parameters (gender or Specification). The best result will be taken depending on the information gain. The best parameter which affects a student's result was the Admission-Type because it has the biggest information gain (0.0636) when general Admission-Type was taken with the gender from the result, the female has the affected parameter, which has information gain of (0.8631), and (0.8113) for male. When state Admission-Type was taken with gender from the result, it is evident that the male is the effecting parameter, with an information gain of (0.4395) and (0.2354) for female. When special Admission-Type was taken with the gender from the result it is evident that female is the affecting parameter. The female was selected because there was only one student in the private admission type.

CONCLUSIONS

The proposed algorithm was applied to the data and the result was significant. ID3 algorithm is an example of symbolic learning and rule induction. It is also a supervised learner i.e. it looks at examples like a training data set to make its decisions. It was developed by J. Ross Quinlan. It is a decision tree that is based on mathematical calculations. A decision tree classifies data using its attributes. The tree has decision nodes and leaf nodes. An attribute is a decision node and a leaf node. ID3 algorithm builds similar decision trees until all the leaf nodes are homogenous. The result shows that the parameter, which affects the students result and makes the right decisions was when general Admission-Type taken together with gender. From the result it could be seen that female is the affecting parameter, when state Admission-Type was taken with the specialization from the result it is evident that the statistics is the affecting parameter. When private Admission-Type was taken with the gender from the result it could be seen female is the affecting parameter. Faculty of Mathematical and Computer Sciences- University of Gezira districts is starting to adopt such an institution-level analysis for detecting areas of improvement, setting policies, and measuring results.

REFERENCES

Andrew Colin, Dr. Dobbs (1996), "Building Decision Trees with the ID3 Algorithm", Journal.

Chen, M.-S., J. Han, *et al*. (1996). "Data mining: an overview from a database perspective." Knowledge and data Engineering, IEEE Transactions on 8(6): 866-883

Crist´obal Romero, Sebasti´an Ventura (2010). "Educational Data Mining: A Review of the State of the Art " VOL. 40, NO. 6.

Fayyad, U., G. Piatesky-Shapiro, (1991). "Data Mining w badaniach rynkowychi merketingowych."

Fayyad, U., G. Piatetsky-Shapiro (1996). "From data mining to knowledge discovery in databases." AI magazine 17(3): 37.

Hem Jyotsana Parashar, S. V., and Nisha Vasudeva (2012). "An Efficient Classification Approach for Data Mining." International Journal of Machine Learning and Computing.

Mohammed M. Abu Tair, A. M. E.-H (2012)," Mining Educational Data to Improve Students' Performance: A Case Study " International Journal of Information and Communication Technology Research (ICT) Journal.

Muslihah Wook, Z. M. Y., and Mohd Zakree Ahmad Nazri (2013). "Preliminary Overview of Data Mining Technology for Knowledge Management System in Institutions of Higher Learning." World Academy of Science, Engineering and Technology.

R. Sumitha , E.S. Vinothkumar (2016). Prediction of Students Outcome Using Data Mining Techniques, (IJSEAS) – Volume-2, Issue-6, ISSN: 2395-3470.

S.Linoff, M. J. A. B. a. G. (2004). "Data Mining Techniques Second Edition."

Sato, Y. (2000). "Perspective on data mining from statistical viewpoints." Knowledge Discovery and Data.

Tom Mitchell (1997) ."Machine Learning", McGraw-Hill, pp. 52-81.