Data Mining: Classification Techniques of Students' Database A Case Study of the Nile Valley University, North Sudan

Tariq O. Fadl Elsid
Dept. Computer Science
Faculty of Science & Technologies - Nile Valley University
Atbara, Sudan

Mirghani. A. Eltahir
Dept. Information Technology
Faculty of Science & Technologies - Nile Valley University
Atbara, Sudan

Abstract

The growth of internet is increasing rapidly and the use of systems become very common. A common main problem that faces any system administration or any users is data increasing per-second, which is stored in different type and format in the servers, learning about students from a huge amount of data including personal details, registration details, evaluation assessment, performance profiles, and many more for students and lecturers alike. Graduation and academic information in the future and maintaining structure and content of the courses according to their previous results become importance. The paper objectives are extract knowledge from incomplete data structure and what the suitable method or technique of data mining to extract knowledge from a huge amount of data about students to help the administration using technology to make a quick decision. Data mining aims to discover useful information or knowledge by using one of data mining techniques, this paper used classification technique to discover knowledge from student's server database, where all students' information were registered and stored. The classification task is used, the classifier tree C4.5, to predict the final academic results, grades, of students. We use classifier tree C4.5 as the method to classify the grades for the students .The data include four years period [2006-2009]. Experiment results show that classification process succeeded in training set. Thus, the predicted instances is similar to the training set, this proves the suggested classification model. Also the efficiency and effectiveness of C4.5 algorithm in predicting the academic results, grades, classification is very good. The model also can improve the efficiency of the academic results retrieving and evidently promote retrieval precision.

Keywords : Data Mining, Classification, Knowledge Discovery in Database (KDD), J48 Algorithm, Weka.

1. Introduction

The advent of information technology in various fields has lead the large volumes of data storage in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge Discovery in Databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [4]. The main functions of data mining are applying various methods and algorithms

in order to discover and extract patterns of stored data [2]. The main objective of this paper is to use data mining methodologies to study student's performance in end General appreciation. Data mining provides many tasks that could be used to study the student performance. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here.

This paper is organized as follows: Section II describes the related work. Section III describes the data mining techniques adopted . Section IV presents results and discussions. Section V concludes the paper. Section VI concludes the paper.

II. RELATED WORKS

Data mining has been widely applied in the higher education field as universities provide huge masses of data. Some of the application is to study factors that affect student retention through monitoring the academic behavior and providing powerful strategies to intervene as proposed by[8].

Bassil [2] proposed a model for typical university information system that based on transforming an operational database whose data are extracted from an already existing operational database. The purpose of the proposed design is to help decision makers and university principles.

Romero and Ventura in [8] introduced a survey of the specific application of data mining in learning management systems and a case study tutorial with the Moodle system.

Ahmed et al. [1] used the classification task to predict the final grade, by presenting a study, that can help the student's instructors to improve the student's performance, by identifying those students who needed special attention to reduce failing and taking appropriate action at right time.

III. Data Mining

Data mining is the process of discovering interesting knowledge from large amount of data stored in databases, database warehouse or other information repositories [5].

Data mining (DM) is a computer-based information system keen to scan massive data repositories, generate information, and discover knowledge. The meaning of the traditional mining term biases the DM in grounds. But, instead of searching natural minerals, the target is knowledge.

DM pursues to find out data patterns, organize information of hidden relationships, structure association rules, estimate unknown items' values to classify objects, compose clusters of homogenous objects, and unveil many kinds of findings that are not easily produced by a classic CBIS (Center for Biotechnology and Interdisciplinary Studies). Thereby, DM outcomes represent a valuable support for decisions-making [7].

DM focuses on knowledge discovery, decision making and recommendation. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered that the Web usage mining is very much involved in educational area.

Clustering

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, marketing, medical diagnostics, computational biology, and many others. Different usage of clustering technique are illustrated in [7][5][8][3].

Classification

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. Classification tries to discover relationships between the attributes that would make it possible to predict the outcome. [6][5] Use Classification technique.

Association Rules

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. [7] Uses this method.

I. Data Mining Process

Data mining is a process to exploration data with different data sources type and tools used. Figure (1) Data mining process is a series of steps which can be summarized in the following:

A. Data collection

Within this stage, target data from source is collected from database server.

	A	B	C
25	2009	CS	good
26	2009	CS	good
27	2009	CS	good
28	2009	CS	good
29	2009	CS	good
30	2009	CS	good
31	2009	CS	good
32	2009	CS	good
33	2009	CS	good
34	2009	CS	pass
35	2009	CS	pass
36	2009	IT	V.good
37	2009	IT	V.good
38	2009	IT	V.good
39	2009	IT	V.good
40	2009	IT	V.good
41	2009	IT	V.good
42	2009	IT	V.good
43	2009	IT	good
44	2009	IT	good
45	2009	IT	good
46	2009	IT	good
47	2009	IT	good
48	2009	IT	good

Figure (1) an example of raw Data

B. Data Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. This is the stage where data are cleaned from noise, their inconsistencies are resolved, and they are integrated and consolidated, in order to be used as input to the next stage of pattern discovery. The techniques that are used here can provide client data elaboration [3]. The different tasks of data preprocessing are:

Data Cleaning: The first step in data preprocessing is to clean the raw data. During this step the available data are examined and irrelevant or redundant items are removed from the dataset. Irrelevant records are deleted during data cleansing.

C. Pattern Discovery

In this stage, knowledge is discovered by classify students according to their graduation degrees. The goal of classification is to identify the distinguishing characteristics of predefined classes, based on a set of instances, e.g. students, of each class [13]. Classification is the technique to map a data item into one of several predefined classes. This requires extraction and selection of features that best describes its properties of a given class or category [4].

D. Pattern Analysis

This is the final step in the data mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used. Relational query languages (SQL) allow users to pose queries for data retrieval. High-level data mining query languages need to be developed to allow users to describe data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered pattern. Pattern analysis able us to do the automatic detection of patterns in data from the same source and make predictions of new data coming from the same source [2].

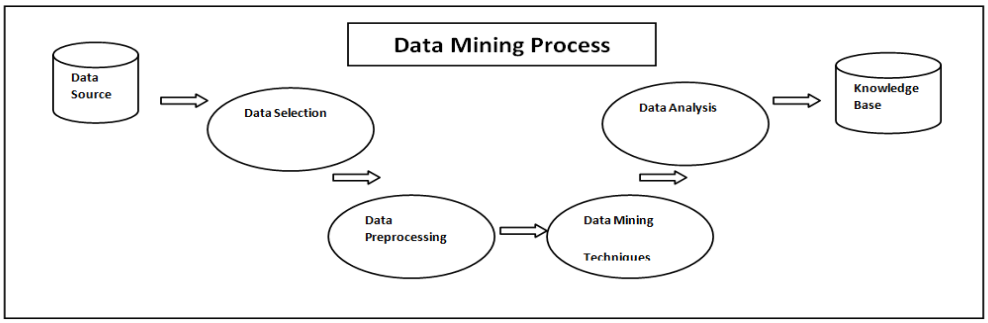


Figure 2: Data mining Process model

Knowledge Post-Processing

In this last stage, the extracted knowledge is evaluated and usually presented in a form that is understandable to humans, e.g. using reports, or visualization techniques. Discovered knowledge must be presented in a user-readable way, using either visualization techniques or natural language constructs. One typical electronic commerce scenario is to provide a marketing expert with patterns s/he can interpret and verify which have to be in the form of rules or visuals. Another is to use the output as input for online prediction and dynamic creation of web pages, depending on the outcome of the built model. A further aspect of knowledge post-processing is knowledge validation. Knowledge must be validated before it can be used for marketing actions on the Internet. Setting up a knowledge maintenance mechanism consists of re-applying the already set up process for the particular problem or using an incremental methodology that updates the knowledge as the data logged changes [7].

Refinement Process

The examination of the knowledge by experts may lead to the refinement process, during which the domain knowledge as well as the actual goal posts of the discovery may be modified. Refinement can take the form of redefining the Internet and marketing data used in the discovery, a change in the methodology used, the user defining additional constraints on the algorithm(s), modification of the marketing knowledge used or calibration of parameters. Once the refinement is completed, the pattern discovery and knowledge post-processing stages are repeated. Note that the refinement process is not a stage of the data mining process. Instead it constitutes its iterative aspects and may make use of the initial stages of the process that is data prospecting, methodology identification, domain knowledge elicitation, and data pre-processing.

Classification Model Implementation

The model has been performed where the students' academic results data collected over a period of time at the Faculty of Science & Technology, Nile Valley University. The main aim of the study was to find how our model managed to classify the new instances(tested set) on supplied instances(training set), we use the final results of students computer science and information technology departments for the year 2006 to 2009), Faculty of Science & Technology.

The model to classify the instances of the Table 1(sample of the train set), the file ARFF(data.arff),WEKA source file which include final grade of graduated students , is depicted in Figure 3 . The predicted instances(tested set), the file ARFF(datanew.arff) see Figure 4, note that the attribute section is identical to the training data, the file includes the tested instances of the attribute ("grade"). The values of "grade" attribute is left as "?", thus WEKA has no actual values to which it can compare the predicted values of new instances.

Table 1: Sample of students' academic results ,their department(CS,IT) and admission year

grade	Deprt(Department)	adm_year
VG (Very good)	CS (Computer Science)	2009
VG (Very good)	CS (Computer Science)	2009
VG (Very good)	CS (Computer Science)	2009
good	IT (Information technology)	2008
good	IT (Information technology)	2008

Once we have learned a model, it can be used to classify new unseen data. These notes describe the process of doing graphically by WEKA.

In this model, Classification via Decision Trees in WEKA, we used the the final academic results data to classify new instances using the C4.5 algorithm (the C4.5 is implemented in WEKA by the classifier class: weka.classifiers.trees.J48),see Figure 3. The outcomes of classification processing to train the final results data (data.arff), appear in Figure 7, whereas Figure 6 shows the processing to classify the new instances along with their predicted class value resulting from the application of the model see Figure 8.

The result of the classifier, the model, in graphical versions of the tree appeared in (see Figure 9 and Figure 10).

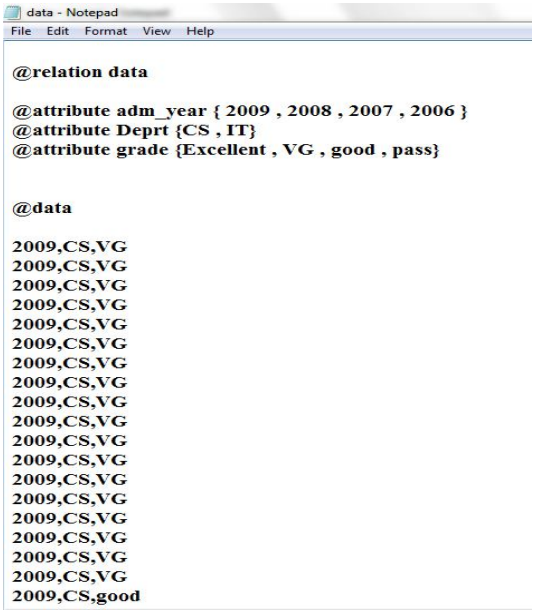


Figure 3: The instances(training set) ARFF file

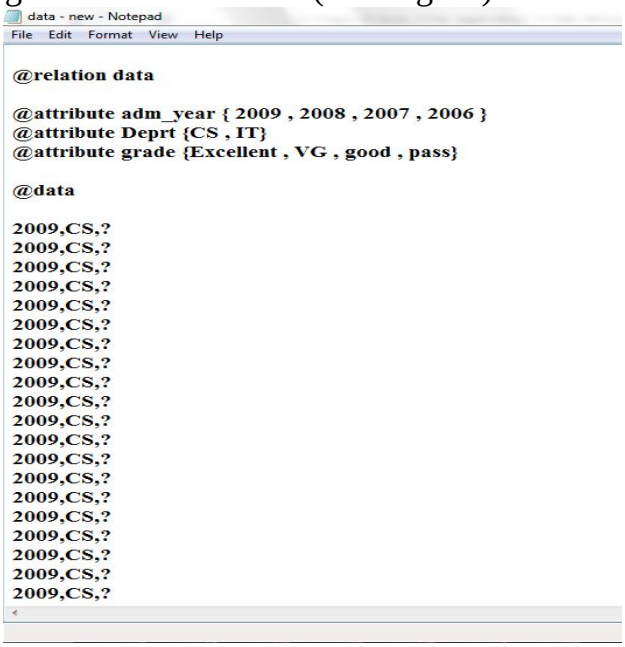


Figure4: The tested instances (tested set) ARFF file

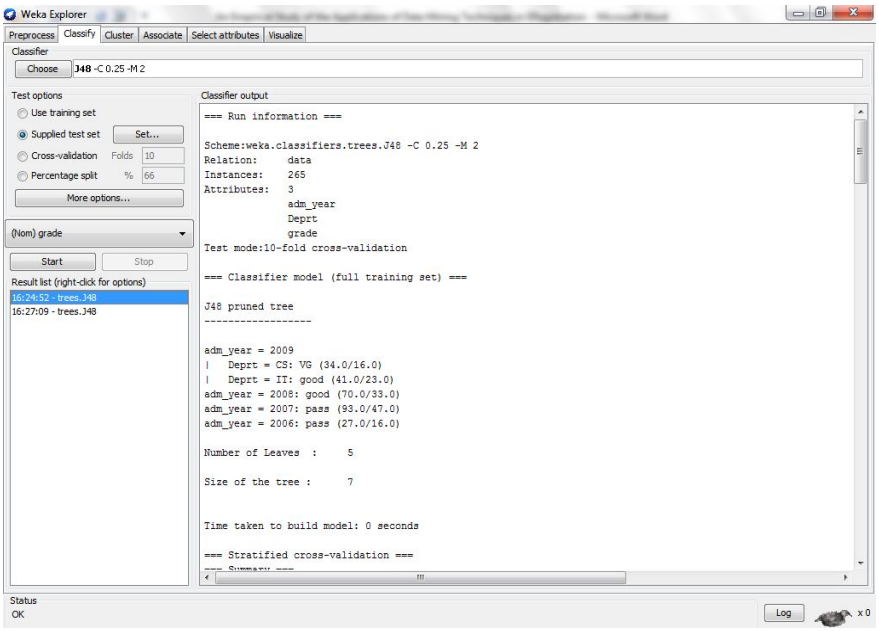


Figure5 : Training Set screenshot of applying C4.5 (J48) classifier, WEKA

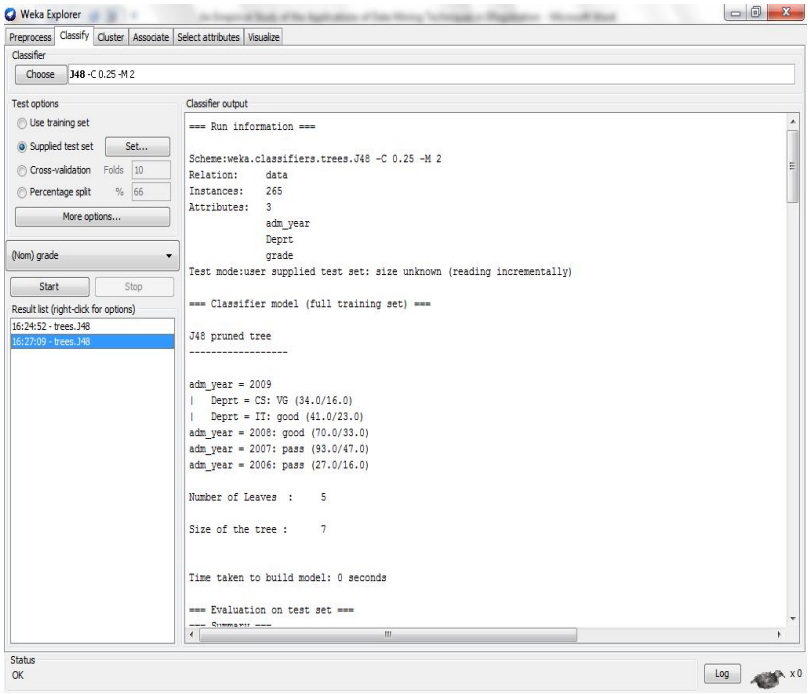


Figure 6: Test set, screenshot of applying C4.5 (J48) classifier, WEKA

Classifier model (training set)

```
 ===Run information===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     data
Instances:    265
Attributes:   3
              adm_year
              Deprt
              grade
Test mode:10-fold cross-validation

 ===Classifier model (full training set)===

J48 pruned tree
------------------

adm_year = 2009
  |Deprt = CS: VG (34.0/16.0)
  |Deprt = IT: good (41.0/23.0)
adm_year = 2008: good (70.0/33.0)
adm_year = 2007: pass (93.0/47.0)
adm_year = 2006: pass (27.0/16.0)

Number of Leaves  :  5

Size of the tree :  7

Time taken to build model: 0 seconds

 ===Stratified cross-validation===
 ===Summary===

Correctly Classified Instances         118               44.5283%
Incorrectly Classified Instances       147               55.4717%
Kappa statistic                          0.1484
Mean absolute error                      0.3143
Root mean squared error                  0.4004
Relative absolute error                 94.477%
Root relative squared error             98.2753%
Total Number of Instances              265

 ===Detailed Accuracy By Class===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.343     0         0           0       0           0         Excellent
                 0.49      0.288     0.205       0.484   0.083       0.205     VG
                 0.588     0.463     0.47        0.456   0.339       0.47      good
                 0.627     0.505     0.615       0.427   0.431       0.615     pass
Weighted Avg.    0.445     0.299     0.452       0.445   0.427       0.574

 ===Confusion Matrix===
```

```
  a   b    c   d     <-- classified as
| 1   0    0   0      a = Excellent
| 35 23 15  0       b = VG
| 39 47 14  0       c = good
  0  2    33 56 |    d = pass
```

Figure 7 : Run information , training set

Test mode : user supplied test set: size unknown (reading incrementally)

```
 ===Run information===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     data
Instances:    265
Attributes:   3
              adm_year
              Deprt
              grade
Test mode:user supplied test set: size unknown (reading incrementally)

 ===Classifier model (full training set)===

J48 pruned tree
------------------

adm_year = 2009
  |Deprt = CS: VG (34.0/16.0)
  |Deprt = IT: good (41.0/23.0)
adm_year = 2008: good (70.0/33.0)
adm_year = 2007: pass (93.0/47.0)
adm_year = 2006: pass (27.0/16.0)

Number of Leaves  :  5

Size of the tree :  7


Time taken to build model: 0 seconds

 ===Evaluation on test set===
 ===Summary===

Total Number of Instances                0
Ignored Class Unknown Instances                265

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

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
?	0	0	0	0	0		Excellent
?	0	0	0	0	0		VG
?	0	0	0	0	0		good
?	0	0	0	0	0		pass
Weighted Avg.	NaN	NaN	NaN	NaN	NaN	NaN	

===Confusion Matrix===

```
 a b c d   <-- classified as
| 0 0 0 0 a = Excellent
| 0 0 0 0 b = VG
| 0 0 0 0 c = good
| 0 0 0 0 d = pass
```

Figure 8 : Run information, tested set

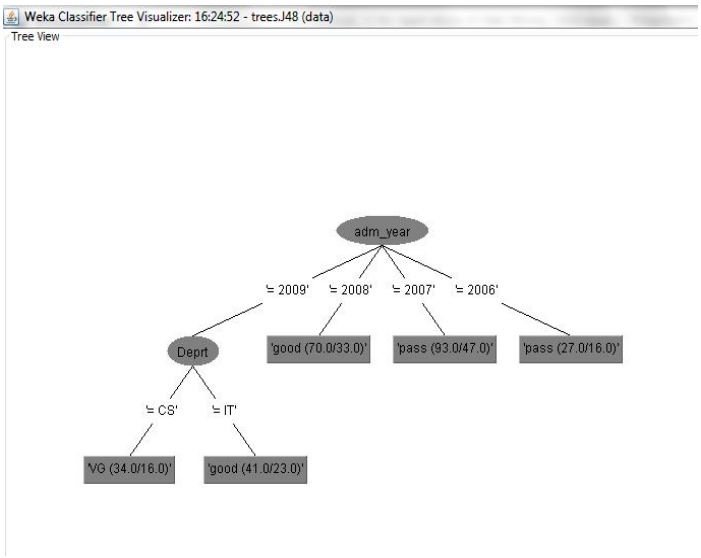


Figure 9 : Decision Tree - for training set

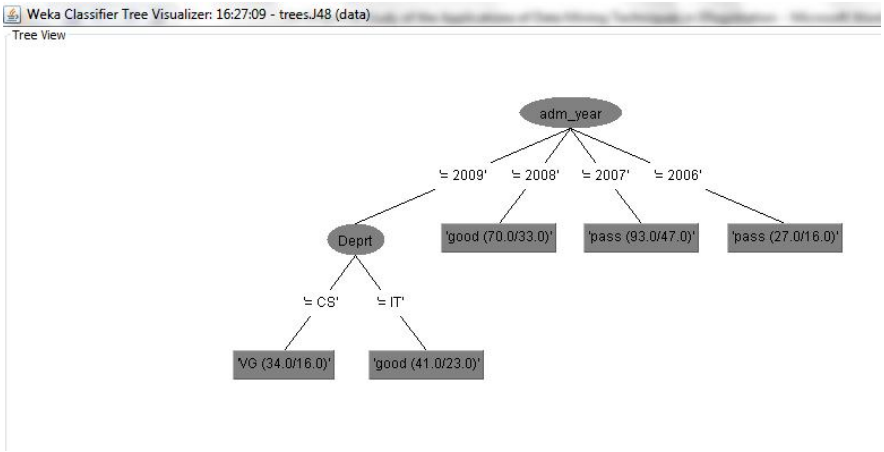
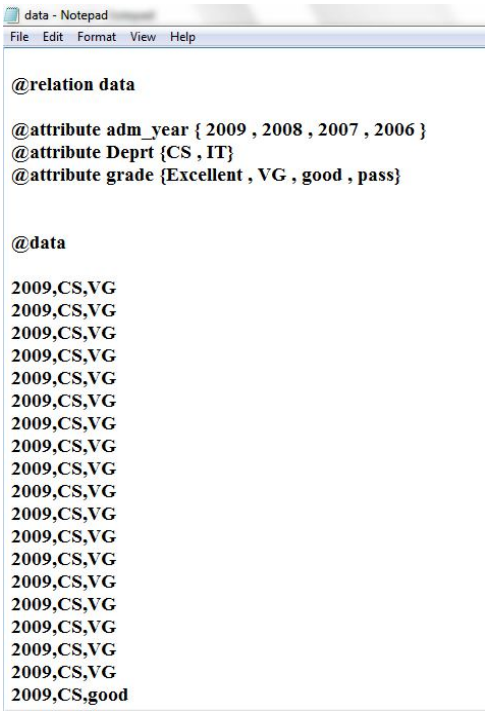


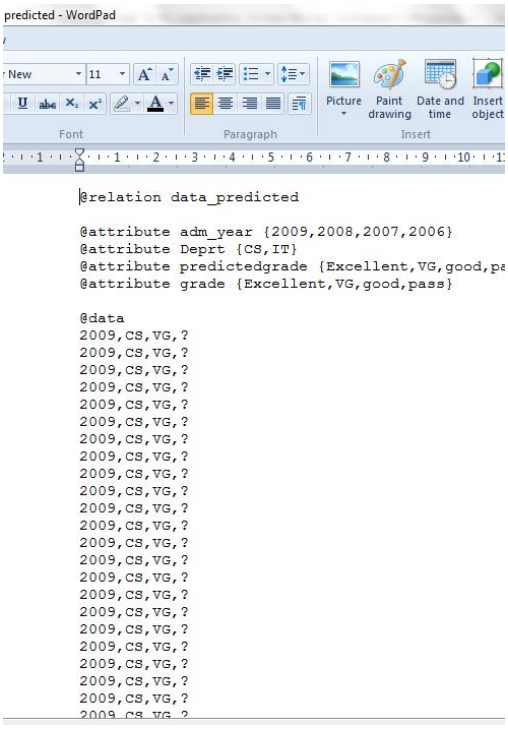
Figure 10 : Predicted Decision Tree - for tested set

IV Results and Discussions

Experiment results show that the efficiency and effectiveness of C4.5 algorithm in predicting the academic results, grades, classification, see Figure 11which shows samples of the classification outcomes for training set and tested set respectively.



(a) training set



(b) tested set

Figure 11: (a) Training set, (b) Tested set

The model also can improve the efficiency of the academic results retrieving and evidently promote retrieval precision. The classification model shows statistical accuracy see Figure 11.

Table 2 : shows statistical accuracies

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0	0	0.343	Excellent
	0.205	0.083	0.484	0.205	0.288	0.49	VG
	0.47	0.339	0.456	0.47	0.463	0.588	good
	0.615	0.431	0.427	0.615	0.505	0.627	pass
Weighted Avg.	0.445	0.299	0.452	0.445	0.427	0.574	

In Table 2 the ROC Area measurement is approximately equals to 0.5 for the classes Excellent, Vg, good, pass, respectively that means the classification process succeeded in training the set. Thus, the predicted instances is similar to the training set , this proves the suggested classification model. The C4.5 algorithm can be used extracting and retrieving information to appear unseen information. The extracted information can be used to achieve the quality in many fields.

V Conclusion and Future work

In the present study, we have discussed the various data mining techniques which can support education system via generating strategic information. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of academic results of students, to solve the problem of failing in academic results for the students in higher institutions. Also, we can use data mining applications in analyzing the students' academic results for long periods, to improve the results. Data mining can be used in the process of choosing applications to the posts by reducing gap between the number of candidates applied for the post, number of applicants .

The algorithms of classification can be as a post-processing step in retrieval information. The feature selection of attributes from large data set is a problem for our future work. For classifications algorithms and others data mining techniques. Communities need to address a number of issues to improve our understanding of data mining, classification.

References

Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. "Data Mining: A prediction for Student's Performance Using Classification Method." World Journal of Computer Application and Technology 2.2 (2014): 43-47.

Bassil, Youssef. "A Data Warehouse Design for A Typical University Information System." arXiv preprint arXiv:1212.2071 (2012).

Erdoğan, Şenol Zafer, and Mehpare Timor. "A data mining application in a student database." Journal of aeronautics and space technologies 2.2 (2005): 53-57.

Jasser, Muhammed Basheer, et al. "Mining Students' Characteristics and Effects on University Preference Choice: A Case Study of Applied Marketing in Higher Education." International Journal of Computer Applications 67.21 (2013): 1-5.

Jiawei Ha, Micheline Kamber. " Data Mining: Concepts and Techniques." Morgan Kaufmann Publishers,2000.

Kantardzic, Mehmed, "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, 2003.

Kumar, Varun, and Anupama Chadha. "An empirical study of the applications of data mining techniques in higher education." International Journal of Advanced Computer Science and Applications 2.3 (2011).

Yu, Chong Ho, et al. "A data mining approach for identifying predictors of student retention from sophomore to junior year." Journal of Data Science 8.2 (2010): 307-325.