Discovering Vital Patterns from UST Students Data by Applying Data Mining Techniques

[1]Ali N. Nusari, [2]Asma A. Al-Shargabi
[1]Faculty of Engineering
Sana'a University
Sana'a, Yemen
anosary@yemen.net.ye

[2]Computer Science Department
Faculty of Science & Engineering, UST
Sana'a, Yemen
somyust@yahoo.com

ABSTRACT: *This paper presents an applied study in data mining and knowledge discovery. It aims at discovering patterns within historical students' academic and financial data at UST (University of Science and Technology) from the year 1993 to 2005 in order to contribute improving academic performance at UST. Results show that these rules concentrate on three main issues, students' academic achievements (successes and failures), students' drop out, and students' financial behavior. Clustering (by K-means algorithm), association rules (by Apriori algorithm) and decision trees by (J48 and Id3 algorithms) techniques have been used to build the data model. Results have been discussed and analyzed comprehensively and then well evaluated by experts in terms of some criteria such as validity, reality, utility, and originality. In addition, practical evaluation using SQL queries have been applied to test the accuracy of produced model (rules).*

Keywords: Data Mining (DM), Knowledge Discovery, Decision Trees, Clustering, Association rules

Received: 7 March 2010, Revised 27 March 2010, Accepted 3 April 2010

©2010 DLINE. All rights reserved

1. Introduction

Currently, Data mining becomes a hot research field which attracts a significant concern. It converts raw data into knowledge that has a meaning for the real world by mapping low-level data into other forms that might be compact, more abstract or more useful. Many successful data mining applications appear and have been deployed in operational use on large-scale real world problems in science and business. Data mining techniques are using today to increase revenues and to reduce costs. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. It is applied in astronomy, finance, fraud detection, manufacturing, telecommunications and Internet agents [1].

On the other hand, higher educational institutions face big challenges in predicting paths of students and alumni. Institutions would like to know, for example, which students will enroll in particular course programs, which students will need assistance in

order to graduate, and are some students more likely to transfer than others. Higher educational institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as drop out, retention, and course success and other academic quality measures in general [2][3].

To benefit from the advantages mentioned earlier for data mining in higher education, this paper will conduct an analyzed case study on UST students data based on data mining techniques. It aims at uncover hidden patterns within data which are modeling students behavior corresponding of academic achievement, demographic students data and financial students behavior.

This paper presents a basic knowledge about data mining and its techniques and also about relevant needs in higher education. It shows literatures and related work relevant to the paper topic. Then, it illustrates our scheme in applying data mining techniques upon the data and also provides a concise description of the target data. Finally, findings are objectively evaluated by UST administration according to some criteria.

2. Data Mining Purposes and Techniques

Traditional methods used to map data into knowledge generally depend on manual analysis. However, data are rapidly accumulated in great manner, thus it becomes difficult to analyze data manually. Data mining provides many techniques and computational tools that assist in extracting useful information (knowledge) from huge databases.

The goal of data mining is to produce new knowledge that the user can act upon. It does this by building a model of the real world based on data collected from a variety of sources which may include corporate transactions, customer histories and demographic information, process control data, and relevant external databases such as credit bureau information or weather data. The result of the model building is a description of patterns and relationships in the data that can be confidently used for prediction [4].

A model is a description of the original historical database from which it was built that can be successfully applied to new data in order to make predictions about missing values or to make statements about expected values [5].

Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart. Data warehouse can be considered as an information system with some special attributes. It is a database designed for analytical tasks using data from multiple applications. It supports a relatively small number of users with relatively long interactions. Its usage is read intensive. Its content is periodically updated (mostly additions). It contains current and historical data to provide a historical perspective of information.

There is some real benefit if your data is already part of a data warehouse. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined [5].
Data mining as a scientific field is can be considered as an extent for many other scientific fields. It governments related fields such as artificial intelligence (AI) and statistic for achieving its goals. So, it is very common to find some algorithms and techniques used in AI, applied in data mining application. However, and as mentioned earlier, the main purpose still patterns discovery.

At abstract level, data mining is used for two general purposes, description and prediction. Under these two abstract purposes, different detailed purposes can be addressed. They can be concluded into six comprehensive tasks: classification, estimation, prediction, affinity grouping, clustering, and description [6]. Classification concern with examine the features of a newly presented object and assigning it to one of a predefined set of classes. The classification task is characterized by a well-defined definition of the classes, and a training set consisting of pre-classified examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it. Decision trees and memory-based reasoning are techniques well-suited for classification.

Link analysis is also useful for classification in certain circumstances. The most common algorithms used for decision trees are Id3, J48, C4.5, CART and CHAID. Where classification deals with discrete outcomes, estimation deals with continuous. In practice, estimation is used for classification. Neural networks are well-suited for estimation. Prediction is the same as classification or estimation except that the records are classified according to some prediction future behavior or estimated

future value. In the prediction task the only way to check the accuracy of the classification is to wait and see. Any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known.

The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior. Market basket analysis, memory-based reasoning, decision trees and artificial neural networks are all suitable for use in prediction. Affinity grouping is used to determine which things go together. The prototypical example is to determine what things go together in a shopping cart at the supermarket, hence the term market basket analysis. Affinity grouping is one simple approach to generating rules from data (association rules).

Apriori algorithm and its extensions is the most common algorithm for association rule. Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous sub-grouping or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. The records are grouping together on the basis of self-similarity. Many algorithms can be considered for clustering.

K-means algorithm and its different variations is the most common known and used for clustering. Sometimes the purpose of data mining is simply to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place. The market basket analysis technique for example, is purely descriptive [4][6].
As a result of earlier insight, different techniques/algorithms are available for different purposes. User (data analyst) has to consider carefully the purpose under consideration, the nature of data and advantages and disadvantages of each technique in terms of some technical criteria.

In this paper, authors have concluded some criteria mentioned in different literatures [6][9][10][11][12] which can assist when making trade-off such as implementation complexity, model type purpose (this means that the target model is predictive or descriptive), preprocessing effort (this means all techniques used for data cleaning, transformations and outliers analysis), post-processing effort (because of results of different algorithms are differing in term of simplicity and clarity, they vary in post processing effort needed to make results), processing time (time needed for processing or for training according to some algorithms), generalization degree (how extent we can generalize the produced model), decision making support, scalability (how extent it can deal with large scale data), and finally, common used or not which reflect simplicity of understanding the technique.

Table I shows a summary of simple comparison between the various data mining techniques/algorithms corresponding to most crucial criteria.

TechniqueCriteria	***Association rules***	***Clustering***	***Decision trees***	***Neural networks***
Pre-processing effort	moderate	moderate	low	high
Post-processing effort	low	high	low	high
Model type purpose	description	description	description	prediction
Processing time	large	Large	large	moderate
Generalization degree	poor	poor	poor	good
Decision making support	good	poor	good	poor
Scalability	low	High	low	medium
Common used	yes	Yes	yes	yes

Table1.Summery of the Comparison Between Popular Data Mining Techniques Corresponding to Suggested Criteria

3. Higher Education Systems and DM

Quality in higher education has been placed front on the contemporary agenda in higher education. There are some quality characteristics and measures in higher educational systems that facilitate evaluating the quality of the system.

In order to evaluate different aspects of the higher educational system, different quality models use different indicators. For example, in systematic approach these indicators can be categorized into three groups: educational system input, process and output. Input indicators of educational system are organized as human resource, financial resources, and systematic resources. In respect to higher educational systems process a series of process indicators can be addressed. These are educating methods, researching methods, teaching methods, studying opportunities, qualitative and quantitative educational improvements. The most important output indicators are alumni and graduates. Some of these measures and characteristics can be used in evaluating the implemented data mining system in a higher educational system.

Nowadays, higher education encounters many challenges which prevent them to achieve their quality objectives. Some of these problems come from knowledge gap in higher educational processes. Knowledge gap is the lack of enough and deep knowledge at educational processes such as evaluation, planning, registration, counselling and marketing. Data mining technology can help bridging knowledge gap in higher educational systems. The hidden patterns that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems [4][5].

4. Related Work

Scientific literatures related to data mining in higher education include number of studies based on different data dimensions and different purposes.

Authors in [8] show such studies which based on students' evaluations using decision tree algorithms, quantitative methodologies adopted to mine data from Learning Management Systems (LMS) to establish usage patterns and online learning designs within the various organizational levels operating in the university. It also shows some studies have built data mining models based on clustering techniques. These techniques have been used to detect cheats in online student assessments, surveys and analysis intended for emphasizing the connections between the university and both master degree studies and continuing education through students' behaviour using questionnaires or data provided by database backed by LMS.

In [14], authors have conducted an applied study about monitoring and evaluation of students' academic achievements using data mining techniques. It has developed a methodology by the derivation of performance prediction indicators to deploy a simple student performance assessment and monitoring system within a teaching and learning environment. The methodology mainly focuses on performance monitoring of students' continuous assessment (tests) and examination scores in order to predict their final achievement status upon graduation. Based on various data mining techniques and the application of machine learning processes, rules are derived to enable the classification of students in their predicted classes.

The result of this research indicates that data mining techniques capabilities provide effective monitoring tools for student academic performance with overall 94% success rating. The various reporting tools that this system offers serve mainly to compare changes over time in performances as may be affected by the different rules that are available plus other well chosen variables exposes systematic structures required to improve performance monitoring.

In attempting to find an association between students' academic achievements and continuing their postgraduate studies, ref. [8] has implemented J48 algorithm using an analysis tool applied on data collected from surveys of different specialization students of the researcher faculty. Main purpose of this study was to predict students who tend to continue their education with postgraduate studies (master degree and Ph.D. studies) through decision trees.

From other side, distance on-line learning with its inherent high technical educational infrastructure and huge number of students represent a fruitful potential opportunity of research by studying quality level of web-based learning systems through analysing students' behaviour on these systems applying text and web mining techniques. Ref. [15] and [16] are two examples of such researches. Ref. [15] has presented a study about the investigation of the effectiveness of web-base learning. It has implemented a special site for this study. It was based on questionnaires and then it has mined them using

quantitative methods of web-mining. It has concluded that learners and instructors could coordinate in a more constructive way. Also, it has concluded that the performance of students could be monitored and these feedbacks can enhance and improve students learning styles. Ref. [16] has designed a text mining system for mining students' electronic portfolios; also it has presented a description of its application integration capabilities through web services.

Generally, we can observe that many of earlier studies tend to focus on analyzing students' evaluations and achievements with less attention to students' demographic data and other educational elements, especially different education processes. Furthermore, they have obviously a lack of clear mapping of the significant results into educational procedures or actionable guidelines for relevant educational institutions to involve these findings within the strategic plans of such institutions. This lack may be result because of that the researchers are always technical specialist rather than educators.

What we have conducted in this research is involving students' demographic data within study to investigate if there are correlations between such data and students' achievements. In addition, to realize the results of this research on UST quality plans, we evolve professional educators, some stakeholders and top managers of UST in the evaluation phase of the produced model. Then we have mapped these results together into educational actionable strategies for more grantee benefit results of this study.

5. UST Data Analysis Scheme

In this section, we describe our scheme followed when we have applied different data mining techniques on UST data. Then, we illustrate the nature of data in terms of its sources, volume and different subject dimensions.

Our scheme uses both directed and undirected knowledge discovery. The process followed in this paper covers the following stages which are summarized by Figure 1.

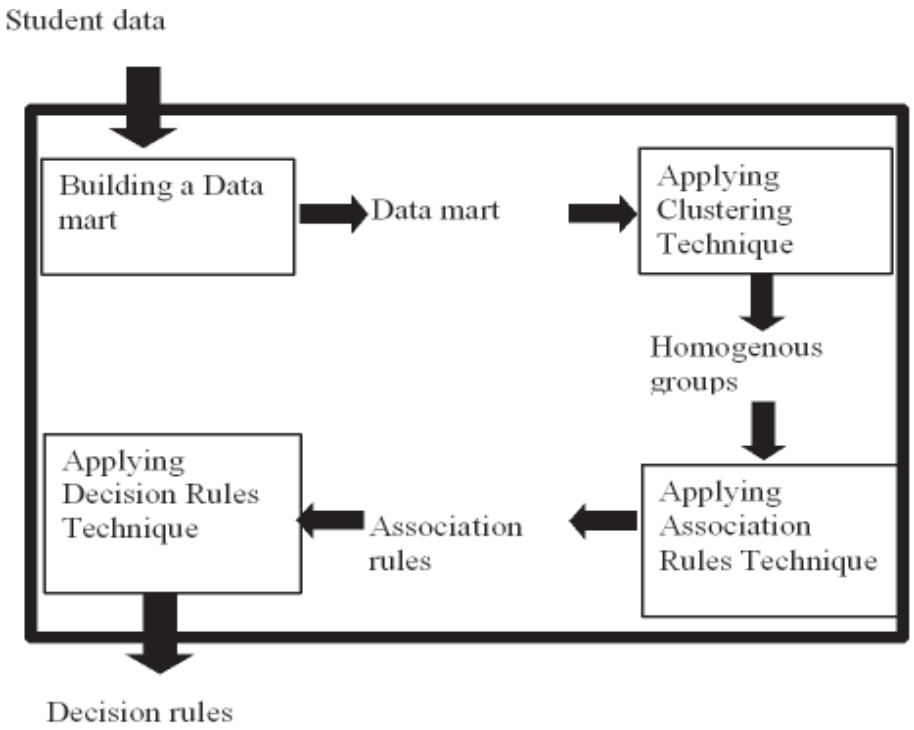

Figure 1. Block diagram of our research scheme

Stage 1 concerns building a reliable data mart (a subset of UST data warehouse which must contain all dimensions of student data). This stage passes through two main phases: databases integration and data preprocessing. In the first phase, concerned databases are integrated from different applications into one and huge database. As a result of this phase, many integration errors are produced. These errors are processed in the second phase by applying preprocessing operations upon the data.

In stage 2, obtaining homogenous groups of data applying clustering technique is achieved. These homogeneous groups present a good description of data and simplify next stages.

Stage 3 concerns determining the relations and patterns of variables by applying association rules algorithms in terms of produced groups at last stage.

In stage 4, we will apply decision trees algorithms to get a Knowledge representation from selected groups of data in driven with obtained association rules. The output of this stage is decision rules.

In stage 5, results are discussed and evaluated in terms of some criteria such as validity, potential utility, simplicity and originality.
Finally, stage 6 suggests some action strategies from the point of finding rules of the research.

As mentioned earlier, data used in this model is UST students' academic and financial data. It contains students' achievements through different academic semesters and contains them financial transactions as well.
One of the most important reasons for choosing this data was the huge volume. Because of that some faculties have changed its names and also some academic programs done after 2005, we find us have constrained to use the data produced before 2005 to maintain the validity of produced patterns. So, we have selected the database of UST that contains records from the academic year 1993 to 2005. Fortunately, it also still huge, it contains about 14,106 students who have 334,211 academic transactions and 55,302 financial transactions.

Source of data is UST registration department database and financial student's affairs department. Database is designed using Oracle SQL. UST database includes 168 tables which contain encoding and operational tables. They also include many tables related to managing Oracle software. Academic data contains personal data such as Id, Name, date of birth, nationality, high school percentage…etc. It also contains data about students' transfer among departments' (internal/external) and students' academic achievement data such as academic accumulated average grade with different courses. Financial affairs data consists all different transactions of students' accounts which are represented as credit and debt vouchers.

6. Applying Data Mining Techniques

This section explains the different stages of implementation. It contains data preprocessing stage and different stages of applying the proposed techniques. For achieving data preprocessing, different Ms-Access queries were applied. In order to apply proposed data mining techniques, WEKA environment has been used. This product has been chosen according to some criteria such as supporting different platforms including various algorithms and wide data input/output options.

First of all, a reliable data mart has been created. For this end, different operations were applied on the original databases. These operations were database integration, missing values filtering, erroneous values processing, variables adaptation and variables decoding. The operations consumed about 80% of the overall time. Most of them have been achieved using different MS-Access queries and WEKA preprocessing operations as well.

The database has been integrated from two Oracle 9i operational databases. The first database belongs to the registration department and the second belongs to financial student affairs. The data has been exported from Oracle into MS-Access format to simplify the other preprocessing operations. Huge database with a lot of garbage data has been produced at this stage, thus many different operations were needed to clean the data and to solve data heterogeneity due to using different data sources.

In this stage, reading tables many times was required to make a precise image about the data. Many interviews with employees in UST have been achieved. These interviews were helpful to make better understanding of the data and to determine which data is not useful. After that, an intensive refinement operation was applied iteratively on database to clean it from waste data. This operation may seem as a trivial operation, but in fact, it has consumed a long time and many operations with the huge database (168 tables, 913 attributes).

In addition, the database has contained different missing values. They fall under variant data types which require different process for each one. These missed values were filled as follows:

- Missed values of student marks and high school certificate percentage were processed using the median value.
- Missed values of high school certificate origin, were processed using the most common value.

Regarding the erroneous values, some mistakes have been observed:

- Some courses names were translated incorrectly from Arabic to English. These names have been retranslated to avoid any misunderstanding with courses names when discovering patterns.
- Within financial transactions, some records have a value zero in the debt balance and a nonzero in credit balance of the student.

In order to deal with this problem, an interview with an accountant of the university, who is working from 1993 till now, has been achieved. She suggested removing these records and thought that those mistakes can appear easily as an error due to the lack of accuracy during data entry. A remove query has been done using MS-Access to remove 1,333 records from 55,302 (2.4%) financial transactions.

Many transformation operations have been applied over different values either to be more appropriate for coming later processing or to be suitable for some data mining techniques used. From an initial population of 334,211 records, a final population of 330,524 records was obtained.

Transformation operations that are applied can be summarized as follows:

- Some values were translated into English such as nationality, course status, and student status.
- Some data were grouped to give more clear understanding for the whole data. For example, student status was grouped into regular and drop out, official status was grouped into regular in pay and not, and academic accumulative average was grouped into grades. This quantization process was achieved using Ms-Access update query as shown by interpretation Tables II, III, IV.
- Remove spaces at all string values as a requirement of most data mining tools. Update SQL query was used to achieve this process.

Finally, encoded values have been decoded by building entity relations between tables in order to merge all tables into one large table with redundant data which is necessary for data mining processing. The encoded attributes were nationality, college and department.

Dept balance	Interpretation
=0	don't pay
< credit balance	don't pay
>= credit balance	pay

Table 2.Pay_Status Interpretation

Average	Interpretation
>=90	excellent
>=80	very good
>=70	good
>=60	accepted
>=50	pass
<50	fail

Table 3. Academic_ Accumulative_ Average Interpretation

From an initial population of 334,211 records, a final population of 330,524 records was obtained. The data mart obtained at this stage consisted of 330,524 (academic student records) x 14 (variables) matrix and 53,969 (financial student records) x 6 (variables) matrix. The variables were classified as follows:

Status	Interpretation
withdraw or aborted	Dropped out
Suspended	suspended
Graduated	graduated
Regular	regular

Table 4. Student_Status Interpretation

- ***Eleven numerical variables which are:***
  student_ID, level, joining_year, course_level, course_term, course_mark, debt_balance, credit_balance, account_type, flag_scholarship, flag_financial_assistant.

- ***Nine category variables which are:***
  faculty, department, course_register_status, student_status, course_name, nationality, academic_accumulated_average, high_school_percentage, pay_status.

As mentioned in the previous section, the first technique was applied is clustering. Clustering was applied to data using different algorithms in WEKA environment. In order to obtain a global optimal solution, Iterative implementations of k-means algorithm have been applied with different initial random values. When applying these algorithms, some data transformations were required. For instance with K-means algorithm, strings variables are not accepted. WEKA provides many transformations that can help in this stage. Thus, string variables were transformed into nominal (numerated string values).

Keeping in mind that Yemeni students represent the most population of data (85%), thus, the (Yemeni) value appeared a lot in clusters. The same reason can be valid for (Regular) students (39%).

As shown in Table V, the instances are grouped into 7 clusters. Clusters 2, 3, 5, and 6 show that grouping students in these clusters tend to have a very poor score (fail). The larger clusters 0 and 3 also indicate poor academic achievement (pass) and (fail).

Attribute / Cluster	Faculty	Status	Accumulated Score	Payment Behavior
0	Science& Engineering	graduate	pass	pay
1	The National College	Dropped out	pass	pay
2	Science &Engineering	Dropped out	fail	don't pay
3	Medical Sciences	regular	fail	pay
4	Arts	remain	pass	pay
5	Arts	regular	fail	pay
6	Administrative Sciences	regular	fail	pay

Table 5. Attributes of Different Clusters

In addition, cluster 2 shows that students in this group tend to have a negative behavior toward fees payment. They also then dropped out from the university.

Generally, produced model (clusters) shows that in consideration of issues relevant to academic field, three regions can lay out data:

student drop out, student academic achievements, and fees payment. Thus, next models have selected homogenous grouped according these three attributes, especially with decision rules.

After applying clustering technique in which relation between variables is going to be clear, association rules with guiding lines from clustering will produce clear relations and patterns.

Association rules technique was applied in WEKA environment as well. Some transformations were also required such as string variables which have been transformed into nominal. Association rules were produced using Apriori algorithm for three times applied to three selected subsets. Concluded features with clustering have been used for selecting subsets.

Seventy two rules have been produced in this stage. According to the three issues discussed early in clustering, some produced rules associated with confidence score confirm some concluded results as shown below:

1. STUD_LEVEL=3 ==> Accum_Avg=Fail conf:(0.81)
2. STUD_STATUS_DESC_L=DROP-OUT ==> Accum_Avg=Fail conf:(0.8)
3. STUD_LEVEL=4 ==> Accum_Avg=Pass ==> STUD_STATUS_DESC_L=GRADUATED conf:(0.74)
4. STUD_STATUS_DESC_L=GRADUATED Accum_Avg=Pass ==> STUD_LEVEL=4 conf:(0.74)
5. SPEC_L_NAME=IslamicStudies ==> STUD_LEVEL=4 conf:(0.72)
6. CER=Good, STUD_LEVEL=4 ==> STUD_STATUS_DESC_L=GRADUATED conf:(0.7)
7. STUD_STATUS_DESC_L=GRADUATED ==> STUD_LEVEL=4 conf:(0.7)
8. CER=Good, STUD_STATUS_DESC_L =GRADUATED ==> STUD_LEVEL=4 conf:(0.69)
9. STUD_LEVEL=2 ==> Accum_Avg=Fail conf:(0.68)
10. STUD_LEVEL=4 ==> STUD_STATUS_DESC_L =GRADUATED conf:(0.68)
11. Accum_Avg=Accepted ==> STUD_STATUS_DESC_L=GRADUATED conf:(0.66)
12. STUD_STATUS_DESC_L=REGULAR ==> Accum_Avg=Fail conf:(0.63)
13. Accum_Avg=Pass ==> STUD_STATUS_DESC_L=GRADUATED conf:(0.62)
14. Accum_Avg=Pass ==> STUD_LEVEL=4 conf:(0.62)
15. CER=Accepted ==> Accum_Avg=Fail conf:(0.55)
16. STUD_STATUS_DESC_L=REGULAR ==> Accum_Avg=Fail conf:(0.63)
17. COLG_L_NAME=Administrative Sciences ==> Accum_Avg=Fail conf:(0.6)
18. COLG_L_NAME=Science & Engineering ==> Accum_Avg=Fail conf:(0.55)
19. STUD_STATUS_DESC_L= DROP-OUT ==> pay=DontPayer conf:(0.61)

As examples, some rules can be concluded from the results as follows:

- Rule 2 with confidence 0.8 clearly shows that dropped out students were failed regardless faculty or department.
- Rule 9 with confidence 0.68 shows that the most students' failures occur at earlier levels, specifically at level 2.
- Students have a poor score at high school; have a poor score in academic accumulative average as shown in rule 15 with confidence 0.55.
- Students that have good scores at high school; will graduate as shown in rule 6 with confidence 0.7 and rule 8 with confidence 0.69.
- Students who pay late or don't pay, usually dropout as shown in rule 19 with confidence 0.61.

Three important rules that were produced (16, 17, and 18) clearly address a remarkable new issue: most regular students do not have a good score in academic accumulative average, especially in Administrative Faculty, and Faculty of Science and Engineering.

After that, decision rules have been obtained with the assistance of results produced in the last two stages. Three classifier attributes (decision attributes) have been selected to apply the technique of decision trees. Classifier attributes were student status, academic accumulated average, and payment status. Id3 and J48 algorithms were used and some data transformations were needed as well. They were also achieved simply using WEKA preprocessing filters.

Training operation in this stage has focused basically on attributes appeared in produced patterns and relations. By tracing produced trees, we can see that many produced patterns in the last stages were confirmed as decisions. Four important decision trees produced seven clear decision rules about students' success and failures, student drop out, and fees payment behavior. The produced model of seven clear decision rules was as follows:

If student_status=Regular then academic_accumelated_average = fail (62%).

If student_status= Dropped-out then academic_accumelated_average = fial (66.6%).

If student_status= Dropped-out and student_level= 1 or 2 or 3 then academic_accumelated_average= fail (60%).

If student_level=1 and academic_accumelated_average=fail then student_status = Dropped-out (90%).

If student_level=2 and academic_accumelated_average=fail then student_status = Dropped-out (60%).

If student_status =DROP OUT then payment_behaviour = don'tpayer (87.5%).

If student_level= 2 or and high_school_score=pass or accepted then student_status = Dropped-out (57%).
We can see that patterns produced by association rules are confirmed here: poor scores in high school affect students' performance in faculty (rule 7). Most regular students have a poor score in accumulated average. Most drop out occurs at earlier levels (rules 3, 7, 4, 5). Student's failures lead to drop out. Students that are dropped out usually pay fees late or don't pay (rule 6). In addition, many produced decision rules confirmed that there are a relationship between students drop out and poor scores in some courses. In general: Most of these courses were at the earlier years (1 or 2). Most of these courses have a complex nature.

7. Discussion and Evaluation

Results have been evaluated by the university academic administration experts and decision makers in consideration of some criteria (validity, reality, utility and originality) on a scale from 1 to 10. These criteria were carefully selected to evaluate the produced model keeping in mind criteria that evaluate the models of decision trees in general. Results have a general score of (8.4). The best evaluated characteristic was the utility with score of (9.5) followed by validity (8.7) and reality (8.5) and finally originality (7).

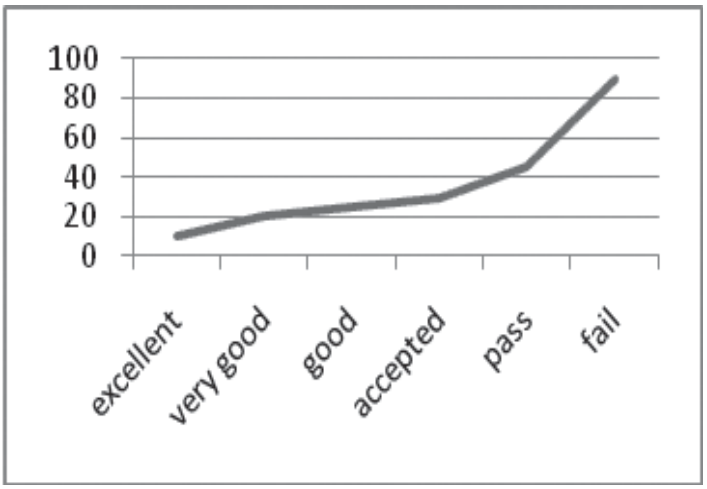

Figure 2. Distribution of student's academic achievements

Keeping in mind all produced results, some of these results were questionable. The common level of achievement for regular students which were poor addresses many inquiries. As shown in Figure 2, the distribution of the academic achievements is so far about the normal distribution in which grade fail represent the most percentage as produced in the model. This result can be justified by various reasons which can be studied by decisions makers at UST.

As shown by results, the most failures are at the early levels, and these failures affect student dropping out, these poor scores may appear due to poor scores at high school, or maybe, students don't know the nature of different department adequately before starting study, so students may surprise with the nature of the department and so, they drop out. Also, some students may have a high ambition not compatible with cognitive abilities of themselves. These types of students may face many disappointments which can be avoided by academic guidance.

Also correct admission policy can help or even adding a primary semester/year to prepare students. Finally, changing positions of some complex courses to upper levels within programs specifications may help reducing the ratio of students' drop out.

Furthermore, produced model shows that dropped out students have a negative behavior regarding to fees payment. Negative feelings which are a result of failures may make students feel not settled and so, they don't easily pay fees.

Also correct admission policy can help or even adding a primary semester/year to prepare students. Finally, changing positions of some complex courses to upper levels within programs specifications may help reducing the ratio of students' drop out.

Furthermore, produced model shows that dropped out students have a negative behavior regarding to fees payment. Negative feelings which are a result of failures may make students feel not settled and so, they don't easily pay fees.

8. Recommended Action Strategies

Finally, results can be mapped into some action strategies:

- Developing an academic guidance program could present a vital service for new students. It could help them determining their abilities and attitudes that will help in choosing the appropriate specialization. Also, this program can present consulting services about study strategies and skills, administrative procedure, psycho-social support for students to support their academic performance.
- Objective admission policy should be designed in each specialization independently in terms of the nature and academic standards.
- Academic programs specification especially courses flowcharts and paths should be carefully reviewed so that each course should be in the appropriate position with its appropriate prerequisites so students can move toward upper level smoothly.
- Some academic programs may in need to a primary semester or even a year to prepare students to study specialization courses later.
- Some courses which have a complex scientific nature which were included in patterns related with failures and dropout may be in need to develop the teaching and learning strategies and methods. It could be taught using tutorials labs to make it easier for students especially if these courses are at earlier levels.
- Supporting students to enhance their academic achievements leads to improvement of fees payment. This crucial pattern can be considered as a strategic policy in UST focusing in supporting students especially at earlier levels.
- Students' achievements evaluations methods of different academic programs should be carefully reviewed and should be strongly involved in academic annual evaluation of university which is responsibility of quality assurance and development department.
- Building a reliable incremental data warehouse involves all branches with different subjects dimensions will help decision makers in UST to improve academic performance.

9. Conclusion

Keeping in mind the aim of this research, the proposed solution has achieved a good result. Findings were well evaluated in context

of some criteria such as validity and reality by the decision makers at UST. Quality assurance departments in higher education institutions could get a useful knowledge about the quality level of the different elements in education systems using such techniques. Discovered patterns could be reformulated into academic strategies about many academic issues as admission policy, students' services, and teaching methods. As farther work, these findings can be improved with expanding dimensions and volume of data such as attendance, students' transfer data, residence data, staff data and other branches data.

References

[1] Luan, Jing (2004). Data Mining Applications in Higher Education.

[2]Fayyad, U. et al (1996).From Data Mining to Knowledge Discovery in databases, American Association for Artificial Intelligence, p. 37-54.

[3]Witten, I.H., Frank, E. (2005). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2nd Edition.

[4]Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery, Scientific Report.

[5]Berson, Alex Smith, Stephen Thearling, Kurt Naeimeh (2000). Building Data Mining Applications for CRM, McGraw Hill.

[6] Berry, Michael et al (1997). Data Mining Techniques- For Marketing, Sales, and Customer Support, Wiley.

[7] Delavaril,Naeimeh et al (2004). A New Model for Using Data Mining Technology in Higher Educational Systems, IEEE.

[8] Paul, Vasile., Bre_felean (2007). Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment, Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28.

[9] Mitchell, Tom., M Linoff, Gordon (1997). Machine Learning, McGraw Hill.

[10] Aijun An, ShakilKhan, Xiangii Huang (2003). Objective and Subjective Algorithms for Grouping Association Rules, *In*: Proceedings of the Third IEEE International Conference on Data Mining (ICDM03) P(1).

[11] Lekas, George Konstantinou (2000). Data Mining The Web: The Case of City University Log's Files.

[12] Bradly, P., Fayyad U., Reina C. (1998). Scaling Clustering Algorithms to Large Databases.

[13] Ramasubramanian, P. et al (2007). Mining Analysis of SIS database using Rough Set Theory, IEEE, International Conference on Computational Intelligence and Multimedia Applications.

[14] Ogor, Emmanuel, N. (2007). Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques, Fourth Congress of Electronics, Robotics and Automotive Mechanics, IEEE.

[15] Lau, Irene Kuan., Fong, Joseph (2003). Investigation on the Effectiveness on Web-Based Learning Using Data Mining Approach, *In*: Proceedings of the 14th International Workshop on Database and Expert Systems Applications, IEEE, 2003.

[16] Garc´ýa Adeva, Juan Jos´e et al (2006). Applying Plagiarism Detection to Engineering Education, IEEE.