DEVELOPMENT OF SOFTWARE SYSTEM FOR DATA MINING OF STUDENTS’ SURVEYS FOR ASSESSMENT OF COURSES TEACHING QUALITY IN ENGINEERING COLLEGE OF BAGHDAD

GRADUATE QUALIFICATION WORK
SUSU- 02.04.02.2018.308-587.GQW

Supervisor
Cand. Sci., Assoc. Prof.

________________________ M.L. Zymbler

Author,
the student of the group CE-219

________________________ A.H. Al-Taleb

Normative control

________________________ O.N. Ivanova

“___” _________ 2018

Chelyabinsk–2018
MINISTRY OF EDUCATION AND SCIENCE OF THE RUSSIAN FEDERATION
Federal State Autonomous Educational Institution of Higher Education
South Ural State University (National Research University)
School of Electrical Engineering and Computer Science
Department of System Programming

APPROVED
Head of the department,
Dr. Sci., Prof.
__________________________ L.B. Sokolinsky
“___” ___________ 2018

TASK
of the master graduate qualification work
for the student of the group CE-219
Al-Taleb Amenah Hasan Flaib
in master direction 02.04.02
“Fundamental Informatics and Information Technologies”
(master program “Database Technologies”)

1. The topic (approved by the order of the rector from 04.04.2018 No. 580)
Development of software system for data mining of students’ surveys for assessment of courses teaching quality in Engineering college of Baghdad.

2. The deadline for the completion of the work: 01.06.2018.

3. The source data for the work


4. The list of the development issues

4.1. Review the papers related to topic data mining techniques and education data mining cases.

4.2. Design quizzes to collect student’s opinion on courses in Engineering college.

4.3. Design the software system and implement for quality evaluation of students courses in Engineering college / Baghdad.

4.4. Perform quizzes collect the data and present useful knowledge on courses extracted by the system.

5. Issuance date of the task: 09.02.2018.

Supervisor
Cand. Sci., Assoc. Prof. M.L. Zymbler

The task is taken to perform A.H. Al-Taleb
# TABLE OF CONTENTS

1. **INTRODUCTION** .................................................. 4

2. **REVIEW OF RELATED WORK** ................................. 8
   1.1. Overview of the data mining techniques .................. 8
   1.2. Papers on educational data mining ....................... 11
   1.3. Modern software to perform data mining. ............... 17

3. **DESIGN OF EDUCATIONAL DATA MINING SYSTEM** ...... 19
   2.1. Basic principles ............................................. 19
   2.2. Designing of structure and behavior of
        the educational mining system ............................. 19
   2.3. Design of quizzes to gather educational data .......... 21
   2.4. Design of web-site that will implement quizzes
        for evaluation process ..................................... 23
   2.5. Design of scenarios of applying data mining
        technique ................................................... 27

4. **IMPLEMENTATION** .............................................. 31

5. **EXPERIMENTS AND DISCUSSION** ............................ 38

6. **CONCLUSIONS** .................................................. 45

7. **REFERENCES** ................................................... 46
INTRODUCTION

Topicality of The Research

Data mining is the process of discovering interesting knowledge, such as patterns, association, changes, anomalies and significant structures, from large amounts of data stored in database, data warehouses, or other information repositories [11]. In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining. Educational data mining (also referred to as “EDM”) is defined as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in [5].

E-learning system provides data that is transformed into information used by teachers in order to improve teaching process. Students also benefit from that information, and also by adapting their learning process. The goal is to understand the learning and teaching in order to improve its quality.

EDM might be used for student modeling and e-learning system evaluation. Teachers group or classify students, discover patterns of misconceptions, etc.

In classroom, teachers observe students behavior and analyze test results. They adapt the instruction according to the provided feedback. Students are different and respond to the changes according to their characteristics, so teacher constantly adapts. Such information and feedback is often missing in e-learning systems.

Since such systems store large amount of data, EDM techniques might provide meaning to the data as basis for adaptation, scheduling, prediction of students’ dropout, or course enrollment.

In general, EDM process consists of three phases: data collection and preparation, data analysis and interpretation of results [18].
Background

Student end-of-course evaluations are widely used by colleges and universities to determine the factors that help success of courses and the effectiveness of their instructors, also to define the factors or attributes are often appear together in good and bad courses.

The success of the assessment depends on how this questionnaire has real, accurate and focused details questions about each semester and the teacher associated with it, this helps the head of department to restructure coursework by giving importance to some of the attributes have obvious impact on the final assessment and was ignored previously.

As well as in terms of final evaluation of the teaching, it gives an opportunity for each instructor to identify his weaknesses from the student’s point of view, while making the head of the department or college to reconsider the appropriate instructor depending on the specialization and experience.

In our research, we took the College of Engineering at the University of Baghdad, which consists of 12 scientific departments, where the focus was on the Department of Civil Engineering, which is the first department opened in the university and includes many engineering branches; construction engineering, engineering bridges and roads, environmental engineering, etc.

The duration of study in this college is four years where each year about 7 semesters have studied and these semesters are divided according to their importance to basic and secondary classes. Therefore, the number of lectures per year is differenced and some of these semesters contain a theoretical and practical part.

We took an assessment of approximately 19 semesters at a rate of 400 students divided by the four grades, and the models were divided between the main and secondary classes, containing a practical lesson or not.

For these students, we noticed a significant decrease in the final grades and the attendance in the lectures is relatively low. Therefore, college manage-
ment needs to Proper understanding on the learning behavior of the students helps the educational programs, to a much more improved level, which can increase the learning capabilities of the student, who followed their educational programs.

Here came the benefit of the use various data mining methodologies in different ways to identify learning pattern of student.

**Research goal and objectives**

The goal of the research is the evaluation of the semester courses and their instructors in the Engineering College. This evaluating is based on many factors and attributes that used as questions, which are answered by the students of that college.

For the reaching this goal we must solve the following objectives.

1. Overview the data mining techniques and papers on educational data mining.

2. Design quizzes to ask students about the courses and instructors, and develop an application for performing quizzes and collecting educational data.

3. Design educational data mining scenarios for understanding of the causes for the problems above.

4. Implement the developed scenarios as Data Mining system for discovering hidden knowledge.

5. Implement the developed scenarios as Data Mining system for discovering hidden knowledge.

6. Apply the developed Data Mining system to the collected data and present the extracted knowledge to the College authorities.

**Outline of The Research**

This thesis consists of the introduction, four chapters, conclusions and a list of references.

The first chapter contains overview of the data mining techniques classification, association Rules and clustering, a literature review on previous works
in educational data mining and review of modern software to perform data mining.

Chapter 2 contains basic principles about design quizzes to gather data and the software that implemented them, also describe the scenarios that applied on processing collected data.

Chapter 3 contains implementation of software system.

Chapter 4 contains experiments and discussion of the developed software system.

Conclusion contains summarizing remarks on results of the thesis. The volume of the thesis is 49 pages. The volume of the list of references is 33 sources.
1. REVIEW OF RELATED WORK

1.1. Overview of the Data Mining Techniques

Classification

Classification is used to classify each item in a set of data into one of pre-defined set of classes or groups. The data analysis task classification is where a model or classifier is constructed to predict categorical labels (the class label attributes).

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data [17]. The objects to be classified are generally represented by records in a database table or a file, and the act of classification consists of adding a new column with a class code of some kind [25]. For example, a classification model could be used for text categorization (e.g., spam filtering), fraud detection, optical character recognition, machine vision (e.g., face detection), natural-language processing (e.g., spoken language understanding), market segmentation (e.g., predict if customer will respond to promotion), bioinformatics (e.g., classify proteins according to their function) [28].

A classification task begins with a data set in which the class assignments are known. For example, a classification technique (or classifier) is a systematic approach to building classification models from an input data set. Examples include decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers.

Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability; i.e., models that accurately predict the class labels of
previously unknown records [23].

**Association Rules**

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [Agrawal et al. 1993]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [33].

Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalog design, and loss-leader analysis etc.

For example, if the customer buys bread then he may also buy butter. If the customer buys laptop then he may also buy memory card.

There are two basic criteria that association rules uses, support and confidence. It identifies the relationships and rules generated by analyzing data for frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time [19].

The Algorithms in association rules mining.

**Apriori.** It searches for frequent itemset browsing the lattice of itemsets in breadth. The database is scanned at each level of lattice. Additionally, Apriori uses a pruning technique based on the properties of the itemsets, which are: If an itemset is frequent, all its sub-sets are frequent and not need to be considered.

**DHP.** DHP algorithm (Direct Haching and Pruning) is an extension of the Apriori algorithm, which use the hashing technique with the attempts to efficiently generate large itemsets and reduces the transaction database size. Any transaction that does not contain any frequent k-itemsets cannot contain any frequent (k+1)-itemsets and such a transaction may be marked or removed.

**FPMaX.** FPMaX (Frequent Maximal Item Set) is an algorithm based on FP Tree. It receives a set of transactional data items from relational data model,
two interesting measures Min Support, Min Confidence and then generates Frequent Item Sets with the help of FPTree. During the process of generating Frequent Item Sets, it uses array based structure than tree structure. Additionally, the FPMax is a variation of the FP-growth method, for mining maximal frequent item sets. Since FPMax is a depth-first algorithm, a frequent item set can be a subset only of an already discovered MFI [29].

Clustering

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but 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 [7].

Cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. However, in other cases, cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points. Whether for understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining [32].

Clustering Algorithms

K-means. This is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids.

Agglomerative Hierarchical Clustering. This clustering approach refers to a collection of closely related clustering techniques that produce a hierarchical clustering by starting with each point as a singleton cluster and then re-
peatedly merging the two closest clusters until a single, all-encompassing cluster remains. Some of these techniques have a natural interpretation in terms of graph-based clustering, while others have an interpretation in terms of a prototype-based approach.

**DBSCAN.** This is a density-based clustering algorithm that produces a partitional clustering, in which the number of clusters is automatically determined by the algorithm. Points in low-density regions are classified as noise and omitted; thus, DBSCAN does not produce a complete clustering [31].

### 1.2. Papers on Educational Data Mining

Using data mining in higher education is a recent research field, there are many works in this area, that is because of its potentials to educational institutes.

**Course evaluation**

In work [27] it is shown how useful the application of data mining techniques in course management systems for online instructors, although it can also be applied together in order to obtain interesting information in a more efficient and faster way.

First, instructors can use visualization techniques to obtain a general view of the students usage data, and for example, if they find something strange or irregular in the plots, then they can obtain more detailed information about these events by viewing statistical values.

Or, if they find some similar groups of students in graphs, then they can apply clustering techniques in order to obtain the exact groups students can be divided into, and these groups can also be used to create a classifier in order to classify students.

The classifier shows what the main characteristics of the students in each group are, and it allows new online students to be classified, finally, the instructors can apply association rule mining to discover if there is any relationship
between these characteristics and other attributes, these rules can not only help to classify students, but also to detect the sources of any incongruous values obtained by the students.

Work [15] presented methodology and results of analyzing a large set of undergraduate course evaluations from an Engineering faculty of a major Canadian university.

Regression analysis in this work revealed similar results to those obtained in previous work (using smaller data sets from other institutions), and new insights into the learning strategies of students nowadays, the effect of lecture times, teaching the same course multiple times, course year and course type.

Also presented a novel information-theoretic study on the distribution of responses to the course evaluation questions, which suggested the reasons why classmates may rate a given course and instructor differently, and discovered that some bad courses are still rated highly by some students.

In work [24] there is applying data mining algorithm of K means Clustering, so the institute administration will be able to make groups of faculty members with different parameters for future use, this helps in comparison also, in turn in decision making.

By applying data mining algorithm of Apriori Association rule, the administration will be able to decide on who all faculty members with different parameters are supporting with minimum value for a particular activity, thus helps in decision making in conducting workshops, seminars or conferences etc.

Work [1] utilized a small data set of student course evaluation answers to provide preliminarily analysis on the feasibility of text mining techniques in analyzing these narrative answers, although only small dataset was used in this study, results proved that text mining is a promising technique to analyze short answer textual information in the course of students.

Evaluation sheets are more efficiently than by simply having to read each
comment individually, by analyzing these responses and calculating the Teaching Evaluation Index (TEI) can transform qualitative responses into quantitative information so that one can gain additional insights to evaluate the value of the course from the students perspective.

**Prediction the performance of students**

Work [13] presents a knowledge discovery applied on demographical data of students, in order to explore the factors having impact on the success of university students, knowledge discovery software, called MUSKUP, has been developed and tested on this data.

With this system, all the tasks involved in the knowledge discovery process are performed collectively, in evaluating student performance, decision tree classification technique has been carried out, the classifications attempt to find out which demographic data is most influential on student GPA.

The scope of study is limited to determining profiles of students whose GPA is equal to 2.0 (which is the minimum GPA required for graduation) or greater and of those students whose GPA is equal to 3.0 or greater (honor degree), thus two classification models were obtained.

In the first model, the types of registration to the university and in the second model, the monthly income of the family were found to be the greatest factors affecting the target.

Work [20] conducts experiments using the K-Means clustering technique on data extracted from the Moodle forums on 46 students of Educational Organization and Public Policy course, of the course Physical Education in a higher education institution.

The results were analyzed by the technique Silhouette Index, in order to identify the best training groups in their respective disciplines.

Finally, the interpretation of the results to identify the behavioral aspects of the students was held, based on three dimensions of interaction: General Purpose Profile VLE, Student-Student Interaction and Student-Teacher Interaction
Bidirectional.

Work [26] compares different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses and developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors.

In this article used real data from seven Moodle courses with Cordoba University students, where applied discretization and rebalance preprocessing techniques on the original numerical data in order to verify if better classifier models are obtained.

Finally, it claims that a classifier model appropriate for educational use has to be both accurate and comprehensible for instructors in order to be of use for decision making.

In work [6] the classification task is used on student database to predict the students division on the basis of previous database. As there are many approaches that are used for data classification, the decision tree method is used here.

Informations like Attendance, Class test, Seminar and Assignment marks were collected from previous database of students, to predict the performance at the end of the semester and to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination.

Work [30] improves performance of graduate students where graduate students data collected from the college of Science and Technology in Khanyounis, the data include fifteen years period [1993-2007]. It discovered association rules and sorted the rules using lift metric.

Then used two classification methods which are Rule Induction and Naive Bayesian classifier to predict the Grade of the graduate student, also clustered the students into groups using K-Means clustering algorithm.
Finally, used outlier detection to detect all outliers in the data, two outlier methods are used which are Distance-based Approach and Density-Based Approach, each one of these tasks can be used to improve the performance of graduate student.

In work [22] four classifiers were used to segregate the students, a combination of multiple classifiers leads to a significant accuracy improvement in all 3 cases, weighing the features and using a genetic algorithm to minimize the error rate.

In cases where the number of features is low, the feature weighting worked much better than feature selection, the successful optimization of student classification in all three cases demonstrates the merits of using the LON-CAPA data to predict the final grade of student based on their features, which are extracted from the homework data.

In this work gather more sample data by combining one course data during several semesters to avoid overfitting in the case of 9-Classes and try to find the paths that students usually choose to solve the different types of the problems from activity log to extract more relevant features after this apply Evolutionary Algorithms to find Association Rules and Dependency among the groups of problems (Mathematical, Optional Response, Numerical, Java Applet, and so forth) of LON-CAPA homework data sets.

In work [4] present text replays, a method for generating labels that can be used to train classifiers of student behavior, this work used this method to label data as to whether students are gaming the system, within 20 intelligent tutor units on Algebra.

Text replays are 2-6 times faster per label than previous methods for generating labels, such as quantitative field observations and screen replays, text replays also give precise predictions of student behavior at multiple grain-sizes, allowing the use of both hierarchical classifiers such as Latent Response Models (LRMs), and non-hierarchical classifiers such as Decision Trees.
Analyzing teaching performance of instructors

In work [21] understands the key factors affecting the teaching performance of the instructors in MIS at Bogazici University.

KDD methodology was followed throughout the study. In modeling step, for the purpose of data reduction and variable selection two widely used data mining techniques; stepwise regression and decision trees (CHAID and CART) were applied.

The stepwise regression results are supported by the decision trees findings, according to the results in this work, the most important factor to explain the teaching performance of the instructors is the instructor attitudes that are primarily measured by the evaluation process.

In work [14] examines the factors associated with the assessment of teaching performance of teachers, good prediction of training course that will be teacher obtain its way to reach the highest level of quality in her/his performance. In this work used data which collected from teachers administrative information collected from the Directorate of Education, west of Gaza, Directorate of Education, east of Gaza, and the Directorate of Education in northern Gaza.

Information on training courses by winning teachers in the past three years from 2010 to 2013 were obtained from the training centers of the former Directorates each center individually, several topics discussed by using questionnaire teacher answering its.

Then applied data mining techniques to discover knowledge like association, classification rules (Decision Tree, Rule Induction, K-NN and Naive Bayesian (Kernel)) to determine ways that can help them to better serve the educational process.

In work [2] there are four different classification techniques decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis are used to build classifier models.
Their performances are compared over a data set composed of responses of students to a real course evaluation questionnaire using accuracy, precision, recall, and specificity performance metrics.

Although all the classifier models show comparably high classification performances, C5.0 classifier is the best with respect to accuracy, precision, and specificity.

Work [3] predicts the instructor performance and investigates the factors that affect achievements of student to improve the education system quality.

Turkey Student Evaluation records dataset is considered and run on different data classifier such as J48 Decision Tree, Multilayer Perception, Naive Bayes, and Sequential Minimal Optimization, comparison of all the four classifiers is conducted to predict the accuracy.

1.3. Modern Software to Perform Data Mining

There are six most used free software tools for general data mining that are available today: RapidMiner, R, Weka, KNIME, Orange, and scikit-learn.

The goal is to provide the interested researcher with all the important pros and cons regarding the use of a particular tool.

A comparison of the implemented algorithms covering all areas of data mining (classification, regression, clustering, associative rules, feature selection, evaluation criteria, visualization, etc.) is provided.

In addition, the tools support for the more advanced and specialized research topics (big data, data streams, text mining, etc.) is outlined, where applicable.

The tools are also compared with respect to the community support, based on the available sources [16].

For this thesis, a KNIME (Konstanz Information Miner) was used, is a general-purpose DM tool based on the Eclipse platform, developed and maintained by the Swiss company KNIME.com AG.
Its development started in 2004 at the University of Konstanz, Germany. KNIME is open-source, though commercial licenses exist for companies requiring professional technical support.

The tool adheres to the visual programming paradigm present in most DM tools, where building blocks are placed on a canvas and connected to obtain a visual program.

In KNIME, these building blocks are called nodes, and according to the official website, more than 1,000 nodes are available through the core installation and various extensions.

Nodes are organized in a hierarchy and can be searched by name within an intuitive interface. Each node is documented in detail, and the documentation is automatically shown within the interface once the node is selected.

A large repository of example workflows is available to facilitate quicker learning of the tool, one of the greatest strengths of KNIME is the integration with Weka and R.

Although extensions have to be installed to enable the integration, the installation itself is trivial, Weka integration enables using almost all the functionality available in Weka as KNIME nodes, while R integration enables running R code as a step in the workflow, opening R views and learning models within R.

Several other interesting free extensions are also available, e.g. JFreeChart extension that enables advanced charting, OpenStreetMap extension that enables working with geographical data, etc.

There are also commercial extensions for more specific functionalities. Overall, KNIME seems to be one of the best choices for a user interested in a purely visual programming paradigm with a need for a large variety of nodes [8].
2. DESIGN OF EDUCATIONAL DATA MINING SYSTEM

2.1. Basic principles

There are many uses of data mining techniques in higher education as mentioned earlier in chapter one. However, I developed an analytically system takes into account a lot of the attributes related to the courses and the instructors to study how these attributes affect student grades at the end of of the semester.

For this evaluation process, I designed a new website with a specific sub-domain of official website of Iraqi government (IQ) and designed a questionnaire form on this website contain obligatory questions where answer is chosen via radio-buttons.

There are about 400 students from Engineering college are involved in this research, they evaluated 19 courses taught by 26 instructors by answering the obligatory questions that mentioned above. I collect the data from these responses and transformed it into appropriate forms for mining. After that, built classification model, extract association rules and clustering the courses and students by using knime program [8].

2.2. Designing of structure and behavior of the educational mining system

Use case diagram

Use case diagrams are one of the five diagrams in the UML for modeling the dynamic aspects of systems (activity diagrams, state chart diagrams, sequence diagrams, and collaboration diagrams are four other kinds of diagrams in the UML for modeling the dynamic aspects of systems). Use case diagrams are central to modeling the behavior of a system, a subsystem, or a class. Each one shows a set of use cases and actors and their relationships [9].

The use case diagram of the system (“Quality Evaluation of Students Courses”) consists of two basic users (“student”, “model designer”), the user (“student”) evaluates the course and instructor in use cases (evaluate), the user (“model designer”) prepare and design the form of questions (questionnaire) in
use case (prepare questionnaire), and obtain the results from this evaluation and process them in use case (collect results), fig. 1.

Fig. 1. Use case diagram

**Deployment diagram**

A deployment diagram is a diagram that shows the configuration of runtime processing nodes and the components that live on them. Deployment diagrams are one of the two kinds of diagrams used in modeling the physical aspects of an object-oriented system. It is used to model the static deployment view of a system (topology of the hardware).

A deployment diagram is just a special kind of class diagram, which focuses on a nodes of system. Graphically, a deployment diagram is a collection of vertices and arcs. Deployment diagrams commonly contain Nodes and Dependency and association relationships. It may also contain notes and constraints. Deployment diagrams are important for visualizing, specifying, and documenting embedded, client/server, and distributed systems and also for managing executable systems through forward and reverse engineering [9].

The student fills the questionnaire published on website that designed for this purpose. So, the results saved in database on server, after this model designer build model in analysis program and upload this database to obtain
results, fig. 2.

![Deployment Diagram](image)

Fig. 2. deployment diagram

### 2.3. Design of Quizzes to Gather Educational Data

The questionnaire was prepared has accurate and focused details questions about each semester and the teacher associated with it.

Table 1 lists 19 courses semester taken at Engineering college with important information for every semester taken from head of of Civil Engineering department: time of lecture (we define morning classes as those which start between 9:00 and 3:00, and evening classes as those which start after 17:00), number of lectures during all the course (one three-hour lecture, two 90-minute lectures per week), and the type of course (essential or not). Also we get the final graded exam for 400 students with attendance of them through every course mentioned in table 1.

Table 2 lists 19 questions on our evaluation form, we will refer to them by their abbreviations (e.g., C1) C1 through C9 refer to course attributes for example labortary work, physical environment,textbooks/handouts, etc. I1 through I10 refer to teaching attributes for example level of explanation, enthusiastic, etc. Each question has five possible answers from 5 (best) to 1 (worst).
### Tab. 1. Courses Information

<table>
<thead>
<tr>
<th>No</th>
<th>Course Name</th>
<th>Time of lecture</th>
<th>Number of lectures</th>
<th>Course type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mathematics</td>
<td>morning/evening class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>2</td>
<td>Engineering Mechanics</td>
<td>morning/evening class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>3</td>
<td>Building Materials</td>
<td>morning/evening class</td>
<td>13</td>
<td>essential</td>
</tr>
<tr>
<td>4</td>
<td>Engineering Drawing</td>
<td>morning/evening class</td>
<td>13</td>
<td>essential</td>
</tr>
<tr>
<td>5</td>
<td>Strength of Materials</td>
<td>morning/evening class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>6</td>
<td>Engineering Surveying</td>
<td>morning/evening class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>7</td>
<td>Concrete Technology</td>
<td>morning/evening class</td>
<td>13</td>
<td>essential</td>
</tr>
<tr>
<td>8</td>
<td>Building Construction</td>
<td>morning/evening class</td>
<td>13</td>
<td>secondary</td>
</tr>
<tr>
<td>9</td>
<td>Theory of Structures</td>
<td>morning class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>10</td>
<td>Reinforced Concrete</td>
<td>morning class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>11</td>
<td>Engineering Analysis</td>
<td>morning class</td>
<td>28</td>
<td>secondary</td>
</tr>
<tr>
<td>12</td>
<td>Traffic Engineering</td>
<td>morning class</td>
<td>13</td>
<td>secondary</td>
</tr>
<tr>
<td>13</td>
<td>Irrigation and Drainage</td>
<td>morning class</td>
<td>28</td>
<td>secondary</td>
</tr>
<tr>
<td>14</td>
<td>Foundation Engineering</td>
<td>morning class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>15</td>
<td>Concrete Structures</td>
<td>morning class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>16</td>
<td>Steel Structures</td>
<td>morning class</td>
<td>28</td>
<td>secondary</td>
</tr>
<tr>
<td>17</td>
<td>Hydraulic Structures</td>
<td>morning class</td>
<td>28</td>
<td>secondary</td>
</tr>
<tr>
<td>18</td>
<td>Sanitary Engineering</td>
<td>morning class</td>
<td>28</td>
<td>essential</td>
</tr>
<tr>
<td>19</td>
<td>Highways Engineering</td>
<td>morning class</td>
<td>14</td>
<td>essential</td>
</tr>
</tbody>
</table>

### Tab. 2. Questions on course evaluation form

| C1  | Were the materials, e.g. textbooks/handouts, etc. prescribed satisfactory and useful? |
| C2  | Did the laboratory work meet the needs of the course? |
| C3  | How well did the course assignments/examinations reflect the content of the course? |
| C4  | How would you rate the physical environment in which you take this class? |
| C5  | Difficulty and greater of workloads? |
| C6  | Compared to other courses you have taken, how much did you have or work for this course? |
| C7  | The assignments were appropriate for the level of this class? |
| C8  | The course provided a mixture of explanation and practice? |
| C9  | Rate how confident you feel about your knowledge on the subject? |
| I1  | Instructor’s organization and clarity? |
| I2  | Instructor’s response to questions? |
| I3  | Instructor’s availability and approachability outside of class? |
| I4  | Instructor’s level of explanation? |
| I5  | Instructor’s encouragement to think independently? |
| I6  | Had a well-prepared course outline (syllabus)? |
| I7  | Would you choose to take another course with the same instructor? |
| I8  | How the instructor’s use of technology effective? |
| I9  | Would the instructor encouraged students to join in discussions? |
| I10 | The instructor appeared enthusiastic and interests? |
2.4. Design of web-site that will implement quizzes for evaluation process

Database schema

A database is a collection of related data. By data, we mean known facts that can be recorded and that have implicit meaning. A database is designed, built, and populated with data for a specific purpose, it has an intended group of users and some preconceived applications in which these users are interested, in other words, a database has some source from which data is derived, some degree of interaction with events in the real world, and an audience that is actively interested in its contents [10].

In this system DBMS MySQL was used, it’s free to download and use, it uses standard SQL and Joomla! CMS easily connects with DBMS MySQL, fig. 3 shows this database.

![Database of website](image)

Fig. 3. Database of website

This database schema of system consists of 1 table named “Mytable”, which contains information about all courses as shows in fig. 4.

Mytable consists of 23 fields. These fields are: Response, Course_to_
evaluate, Studentname, Graded_exam, C1, C2, C3, C4, C5, C6, C7, C8, C9, I1, I2, I3, I4, I5, I6, I7, I8, I9, I10. These fields have different types of data: int, char and varchar, the primary key for this table is “Response”, which is automatically incremented. fig. 5.

![Fig. 5. Mytable structure](image)

**Design of interface**

Design focuses on anticipating what users might need to do and ensuring that the interface has elements that are easy to access, understand and use to facilitate those actions [12].

Joomla 3 was used for designing this website. Joomla what is known as a Content Management System or CMS. It a system that allows the site editors to easily add content, text, images, maps, advertisements, etc. The core installation files for Joomla are open source and free, the installation .zip file can be obtained from joomla.org, the core installation allows to create a web site with basic functionality, to extend the functionality we can install what are called Extensions, these Extensions can be free or obtained at a relatively small cost.

I used free Protostar template for Joomla 3, which has a set number of default module positions as shows in fig. 6, the location for each position within the framework of the page is determined in the index.php file for the Protostar template.
Fig. 6. Position of protostar template

Fig. 7 shows the main page for the website that contains evaluation form, it contains the main menu in the top of the page: “Home”, “Questionnaire”, the right side (position-7) contains log in form for administrator and other users, as well I put simple slide show at the banner position this is for Home page, the other page contains (“Questionnaire”) form where the students performed the evaluation process, I will explain the details of design this form in the next section.

Fig. 7. Website of academic survey
Design of evaluation form

HTML forms create a variety of user interface controls to collect input in a Web page, each of the controls typically has a name and a value, where the name is specified in the HTML and the value comes either from user input or from a default value in the HTML, the entire form is associated with the URL of a program that will process the data, and when the user submits the form (usually by pressing a button), the names and values of the controls are sent to the designated URL as a string of the form.

My evaluation form consists of three pages: “Course Evaluation”, “Instructor Evaluation”, “Finalization”, where the (“Student”) must write his or her name and do submit for this form. Figures 8, 9 show some components of this form with HTML code of them.

Fig. 8. HTML code

Fig. 9. HTML code
2.5. Design of scenarios of applying Data Mining technique

Goal 1

What attributes (quizzes on evaluation form that mentioned in section (3) in this chapter) that impact on students’ course estimation and predict it. Where this estimation is responses of students on question No. 9 on evaluation form.

**Question No. 9:** Rate how confident you feel about your knowledge on the subject.

We processed the gathered data in a way that helps to achieve this scenario as shown in the steps below.

1. We took the students’ answers to Question No. 9 which includes the student evaluation of the course.
2. We extracted the median of students’ responses to question No. 9 for each course.
3. We extracted the median of students’ responses to each question for each course.
4. We extracted the average for graded exam of students for every course.
5. Therefore, we have a table with (name of the course, student success rate, median for each question (19 questions), final student estimation of the course (good or not).
6. After I got the table with a processed data, i built a decision tree; in this tree structures branches represent conjunctions of attributes on evaluation form that lead to students’ course estimation.

Goal 2

What attributes (quizzes on evaluation form that mentioned in section (3) in this chapter) that impact on tutors’ course estimation and predict it. Therefore, we must obtain this estimation by following the steps below.

1. For each student, we calculated the average of his answers to all the questions (C1-I10) then collected them with the number of lectures that have attended during the whole semester (Attendances).
2. We found the average of the value extracted in (first step) for all students studying in this course.

3. Through the final student grade in the course (grated exam), we extracted the number of successful students during the course to find the success rate.

4. The semester may be a major semester in the college or secondary and there are a differences between them in the number of final units, how to choose the instructor and the number of hours of weekly lectures, etc. So for each major semester we added (1) point, and for the second semester we added only (0.5) point.

5. Finally, the tutors’ course estimation was extracted based on the value extracted in the second step (the average of the students’ responses with their attendances) with (the success rate of the students during the course) and add (1) or (0.5) point for the major or secondary course.

   After the tutors’ course estimation was found, we processed the gathered data in a way that helps to achieve this scenario as described in the steps below.

   1. We extracted the median for students’ responses to each question (C1-I10) for each course.

   2. We extracted the average for graded exam of students for every course.

   3. Therefore, we have a table with (name of the course, students success rate, median for each question (19 questions), final tutors’ course estimation of the course (good or not).

   4. After I got the table with a processed data, i built a decision tree; in this tree structures branches represent conjunctions of attributes on evaluation form that lead to tutors’ course estimation.

   **Goal 3**

   What course and/or instructor attributes are often appear together in good and bad courses, respectively?

   1. Based on the tutors’ course estimation, which was done in advance,
we separated the courses into good and bad courses.

2. We took each set of courses (good or bad) and found the median for each attribute (C1-I10).

3. Filtered each attribute and classified it under any degree (1-5). Thus, the final table contain attributes that have similar degrees.

4. After I got the table with a processed data, I built two association rules for good and bad courses; in these strong rules discovering interesting relations between attributes in all databases.

**Goal 4**

What are the typical groups of courses?

In this scenario, we depend on the department of college to take the full information about every course. We used the following determinants for every course.

1) name of course.
2) time of classes.
3) success ratio.
4) have tutorial or not.
5) importance of the course (essential or not).
6) number of units.
7) have practical or not.

After I got the table with a processed data, I Used K-Medoids Clustering to predict the typical groups of courses.

**Goal 5**

What are the typical groups among the students?

In this scenario, we used the following determinants for every student.

1) gender of the student.
2) educational level.
3) graduation degree in course.
4) attendance during the course.
After I got the table with all information, I Used K-Medoids Clustering to predict the typical groups of students.

**Goal 6**

Find the relation between the determinants of the courses like: the time of lecture, attendance of the students, importance of the courses and the final grade of every course.

1. There are three types of timing the lectures for every course:
   TA: when the timing in the morning as well as in evening.
   TB: when the timing in the morning only.
   TC: the timing in the evening only.

2. The final grade of every course: we found the average grade for all students through every course, then we divided this grade to three levels: GA: refers to (80-90); GB: refers to (60-80); GC: refers to (40-60).

3. Attendance of the students: Depends on the number of lecture that the students were attended through all the course.
   AA: where the number of lectures (20-28); AB: where the number of lectures (12-20); AC: where the number of lectures (1-12).

4. Importance of the course:
   IA: essential course; IB: secondary course.

5. Based on the tutors’ course estimation, which was done in advance, we classified each course under any degree (1-5), then, filtered each attribute for every course was mentioned above. Thus the final table contains the course that have the same degrees with their attributes.

6. After we got the table with a processed data, we built association rule; in this rule discovering interesting relations between all attributes of the course.
3. IMPLEMENTATION

Structure of Software System

We used Knime program to build our software system, which provides a graphical workbench for visual assembly and interactive execution of data. It enables easy integration of nodes and allows for interactive exploration of analysis results [8].

Several algorithms in data mining were used to create our model: Decision tree, Association rule and Clustering. These algorithms first analyzes the data that provided, looking for specific types of patterns and find the optimal parameters for creating this mining model, these parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.

Build a decision tree to see what attributes impact tutors’ course estimation and predict it.

We explained in the previous chapter, the steps for processing and analyzing the gathered data to become more suitable to access this goal by implementation of decision tree in Knime program, as shown below.

Extracting tutors’ course estimation

We need estimation of the tutors for every course to become as a class label for classification process. So, we created a formula consists from a sets of realistic and important elements for the students and courses at the same time, after this structured and implemented this formula in Knime program as shown in fig. 10, this give the university or department the ability to evaluate any course at any time in very easy way to get the results as shown in fig. 11, where the degree of final estimation is between 1 to 5 point.

![Image](image_url)

Fig. 10. The results of evaluation for two courses
Build learner and predictor decision tree

According to the previous steps the tutors estimation was extracted, then we used as a class label for a decision tree that was built, as shown in fig. 12, 13.

![Diagram of decision tree](image)

Fig. 11. Structure of the formula in Knime program

![File reader node](image)

Fig. 12. File reader node

Extracting the accuracy of a prediction

We mentioned in a previous chapters the total responses of students are 400 responses, we used 280 responses for learner decision tree, and the rest of responses are used as test data to predict the tutors’ course estimation, so we added (scorer) node to calculating the accuracy of prediction.

Build a decision tree to see what attributes impact students’ course estimation and predict it.

We explained in the previous chapter, the steps for processing and analyzing the gathered data to built decision tree in Knime program for students’
Fig. 13. Structure of the decision tree in Knime program

course estimation. In this scenario, we took estimation of every student to every course and consider it as a class label for decision tree that we build as shown in fig. 14, 15.

Fig. 14. File reader node

Fig. 15. Learner decision tree node
Apply association rules to understand what the relation between the determinants and outputs of the course like final graded exam, attendance.

We explained in the previous chapter, the steps for processing and analyzing the gathered data to built association rule in Knime program, in this rule, we used file reader node to input our data file, where each row represents one course with its attributes items, as shown in fig. 16, then we found frequent itemsets by adding association rule learner node which used minimum support 0.2 point and 0.7 point for confidence of association rule as shown in fig. 17.

![Fig. 16. File reader node](image)

![Fig. 17. Workflow of association rule](image)

Apply association rules to understand what course and/or instructor attributes are often appear together in good and bad courses, respectively.

We explained in the previous chapter, the steps for processing and analyzing the gathered data to built association rule in Knime program, in this rule, we divided the courses to two parts: good and bad, depends on the tutors course estimation that was mentioned in this chapter section 3.1, so row splitter node
was used to split these courses as shown in fig. 18, for every type of course, we build association rule for it as shown in fig. 19. For every type of course, we build association rule for it, the steps to built these rules are similar to steps that mentioned above.

Using K-Medoids Clustering to predict the typical groups of courses.

We explained in the previous chapter, the requirements and determinants that was needed for every course to clustering it, we clustered the courses using K-medoids algorithm, which chooses datapoints or medoids that can be defined as the object of a cluster whose average dissimilarity to all the objects in
the cluster is minimal, and it is a most centrally located point in the cluster.

Fig. 20 shows the workflow of clustering the courses, in which we used file reader node to input our data file, where each row represents one course with its attributes items, as shown in fig. 21, we used numeric distances node that allow the application of various distances measures in combination with the K-medoids node, it measures the numeric distance between objects in data sets, as shown in fig. 22.

Fig. 20. Workflow of K-Medoids Clustering

Fig. 21. File reader node

Fig. 22. K-medoids node
Using K-Medoids Clustering to predict the typical groups among the students.

We used the same steps and algorithm that used in the previous section but the difference here is about clustering 400 student to three clusters depends on some criteria about every student like: gender of the student, educational level, graduation degree in course and attendance during all the course, as shown in fig. 23.

![File reader node](image)

Fig. 23. File reader node
4. EXPERIMENTS AND DISCUSSION

Build a decision tree to see what attributes impact tutors’ course estimation and predict it.

Results

After implementation the workflow of a decision tree in Knime program, we got the results as shown in fig. 24.

Fig. 24. Results of decision tree and the accuracy of prediction

Discussion

As mentioned in section 2.3, there are 19 quizzes filled by students and they considered as an attributes affected on tutors’ course estimation. So, from obtained results the most important and influential attribute is I4 which represents (Instructors’ level of explanation), in fact this is a realistic result because we depend on a very important factors when we obtained the tutors estimation like final graded, attendance, etc. so the ways that the instructor uses for explanation the materials and his or her ability for reaching the information to students consider the first thing effects on students and course at the same time.

And the second most influential attribute is I1 that represents (Instructors’ organization and clarity) which means how the instructor manages his class in
everything for example in preparing his lectures, develops an appropriate questions for quizzes and exams, knowledge the ways of dealing with students in a way encourage them to attend the lectures, also the clarity that the instructor has so he doesn’t depend on a complex ways or words for dealing with their students.

**Build a decision tree to see what attributes impact students’ course estimation and predict it.**

**Results**

After implementation the workflow of a decision tree in Knime program, we got the results as shown in fig. 25.

![Decision Tree Diagram]

**Fig. 25. Results of decision tree and the accuracy of prediction**

**Discussion**

Students’ course estimation depends on the amount of confidence that the students have towards this material and its impact on their evaluation at the end of semester, so as shown in results the most influential attribute on this estimation is I7 which represent Would you choose to take another course with the same instructor?, this attribute is a question answered by 400 students and shows the importance of instructor performance on this estimation, so when the student evaluates the course and material first thing important to him is the in-

39
structor of this material.

The second most influential attribute is I4 that represents (Instructors’ level of explanation), as we see is the same most influential attribute in tutors’ course estimation, means the ability and experience of tutor to explain the materials for students takes the same importance when the tutors evaluate the course or students.

**Apply association rules to understand what the relation between the determinants and outputs of the course, like final graded exam, attendance, etc.**

**Results**

After implementation the workflow of an association rules in Knime program, we got the results as shown in fig. 26.

![Association Rule Learner](image)

**Fig. 26. Results of association rules**

**Discussion**

We got 15 important association rules with high support and confidence, we explained the parameter of these rules in section 2.6, as shown in the results the good attendance of students (AA): 20-28 is when the type of course isn’t essential (IB) due to get the course high credit exam (GA): 80-90 in the end of semester, in most the cases of rules when the type of course is essential (IA) due to bad graduation credit (GC), also in morning classes (TB) there are more attendance of students (AA) so there is a good impact of the time of lectures on the attendance and final credit.

**Apply association rules to understand what course and/or instructor**
attributes are often appear together in good and bad courses, respectively.

Results

After implementation the workflow of an association rules in Knime program, we got the results as shown in fig. 27.

![Interactive Table]

Fig. 27. Results of association rules

Discussion

The results for a good courses association rule represent the relation between the 19 attributes about the course and instructor at the same time and their impact to make the course in a good classification so, the materials, e.g. textbooks/handouts/videos/slides (C1) and the laboratory work (C2) was combined with instructor’s response to questions (I2), instructor’s availability outside of class (I3) and the enthusiastic of instructor (I10) leads to difficulty and greater of workloads (C5), the opinion of students to take another course with the same instructor (I7) impact by difficulty and greater of workloads (C5), instructors organization and clarity (I1) has a good impact on how the assignments were appropriate for the level of this class (C7) then the confident that the student feel about his knowledge on the subject (C9).

In the case of the results of a bad courses association rules, there is a clear attribute appears in the most of the rules is (I4: instructor’s level of explanation) which means this attributes is the most attribute affected by other attributes and
effects on the final estimation of the course, so when the laboratory work meets
the needs of the course and there is a mixture of explanation and practice leads
to good instructor’s level of explanation and this result a high confident of stu-
dents in this materials (C9).

And another important attribute is Would you choose to take another
course with the same instructor? this question is apply to students and there
are many attributes impact on their answers for example: laboratory work, the
course assignments/quizzes/examinations, confident that the student has toward
this course.

**Using K-Medoids Clustering to predict the typical groups of courses.**

**Results**

After implementation the workflow of a K-Medoids Clustering in Knime
program, we got the results as shown in fig. 28, 29.

![Fig. 28. Scatter plot of K-Medoids Clustering](image)

**Discussion**

We clustered the courses into three types depending on many factors men-
tioned in previous chapter, so as shown in the results the first type of clustered
courses has excellent value of final graded exam coincides with presence of the
tutorial part and absence the practical work, this mean the importance of the tu-
torial part where the instructor take additional lectures to solve many questions
and discuss the students if they have any problems in this material.
Fig. 29. Results of clustering courses

Second type of clustered courses has a good final graded exam coincides with a good practical part, the material of these courses contains practical aspect more than the theoretical.

Third type is the more interesting type of clustered courses, which represents the essential materials that have a high number of units of the final credit of course, so these materials are more complex than others and have a low value of success ratio between 40-62 degree coincides with presence of the tutorial part and absence the practical work.

Using K-Medoids Clustering to predict the typical groups among the students.

Results

After implementation the workflow of a K-Medoids Clustering in Knime program, we got the results as shown in fig. 30, 31.

Fig. 30. Scatter plot of K-Medoids Clustering


Discussion

We clustered the students into three types depending on the gender of every students, stage level of them, final graded of them and attendance (the number of lectures that the student attended). First clustering of student with a good final success ratio coincide with a female more than male students and at the first stage of study.

second and third clustering represent students on the three and four stage of college and have a satisfactory success ratio.
CONCLUSIONS

In this research, we presented our methodology and results of analyzing a large set of undergraduate course evaluations from an Engineering faculty. We have shown how useful the application of data mining techniques in course management systems, although we have shown these techniques separately, they can also be applied together in order to obtain interesting information in a more efficient and faster way.

Based on our analysis, in order to improve the teaching quality, instructors should consider enhance their attitude, and organization visual presentation skills. They want to make sure that they respond questions well and clearly. In order to improve the course quality, instructors may want to design tests and assignments such that they are closely related to the course material.

Due to the low evaluations on the usefulness of textbooks and tutorials, instructors may consider improving the quality of textbook and tutorials. From an institution’s perspective, it should try to schedule more morning classes with smaller number of students since evening classes and classes with large size receive worse evaluations.

Also, we applied clustering techniques to obtain the exact groups of students can be divided into. And these groups can also be used to create a classifier in order to classify students. The classifier shows what the main characteristics of the students in each group are, and it allows new online students to be classified. Finally, we applied association rule mining on the most important deterministic of the course and found when the type of the course is essential this mean more theoretical part and less practical aspect with many workload on students, leads to less final success ratio and more absence from the lectures.
REFERENCES


20. Machado C., Jos R., Lima B. An investigation of students behavior


