# Finding Most Effective Sequences of Modules for Students in Higher Education to Improve their Performance using Knowledge Discovery in Databases

Fatiha Tayeb Bousbahi
Information Technology Department
College of Computer and Information Sciences, King Saud University
Riyadh 11543, Saudi Arabia

## ABSTRACT

Universities offer a wide selection of courses to students studying in different fields; however, this has generated various unintended problem such as students struggling, dropping or failing. Universities also, own large amounts of historical data about students containing rich knowledge that can be used to facilitate a broad range of educational research and analytics study in order to enhance student performance. Knowledge discovery in databases (KDD) is the process of finding comprehensible patterns that can be interpreted as useful or interesting knowledge. One component in this process is Data Mining which is the application of machine learning algorithms for extracting patterns from data. In this context, data mining techniques are used to answer educational research questions that highlight students' success in higher education. Several challenges are facing higher education, one of which is finding the most effective sequence of modules for students for each semester. This finding may help students to improve their performance. In this research we investigate dependency between set of courses chosen by students at each level and their performance using a classification technique in order to build a performance prediction model, which is based on previous students' academic records.

The finding of the study furnishes useful information to advisors in educational institution and valuable knowledge to feed analytics based systems and learning recommender systems to assist students selecting future modules to obtain optimum pathways courses.

## General Terms

Education, Knowledge Discovery, Data Mining, Classification.

## Keywords

Student Performance, Sequence of Modules, Knowledge Discovery in Databases, Educational Data Mining, Classification, Prediction.

## 1. INTRODUCTION

Universities offer a wide selection of courses to students studying in different fields; however this has generated various unintended problems for students such as academic delay and stumbling, dropping out the course in the middle of the semester or failure. Prior to the beginning of each semester, the students are facing problems to select a set of courses mandatory or optional to enroll in. They do not know what to choose. They have advisor but, in general the advisors have good information on the courses they taught but not on all courses. The university supplies to students information about available courses, sections, classrooms and teachers but information regarding most effective sequences of modules from previous students' academic itinerary is missing. In the other hand, universities own amount of data about students. For the researcher, this data represents both an opportunity and a challenge. An opportunity, because there will be work to derive insight from this data. A challenge also, as this data poses problems of all kinds [9].

In an excellent article [9] Grumbach et al. define a data: "A data is the basic description of a reality or fact, such as a temperature record, a student's grade on an exam, account status, message, photo, transaction, etc. A data can therefore be very simple and, taken singly, little useful. But cross-checking with other data becomes very interesting". For example, sequence of modules taken by students at each semester of their academic itinerary with their grades can inform us about student performance. However, data on its own has no value unless it is well examined and processed by algorithms such that provided by Knowledge Discovery in databases (KDD). KDD is a set of methods and algorithms that allow data to talk to us, in other terms, KDD makes clear the meaning behind data [2];[7]. And, the greater the amounts of data, the better the learning algorithms are and consequently, the greater the credibility of the knowledge discovered.

Facing the phenomenon of academic delay and stumbling, many academic researchers tackled student performance prediction using Data Mining techniques to identify the causes of these problems **Error! Reference source not found.**.

The present work is an exploratory analysis of students' recorded data in our workplace, the female campus at Information Technology department (IT) at the College of Computer and Information Science (CCIS) at King Saud University (KSU). Each semester called level, the students have to enroll in a set of modules from their department and others. A course is available to a student when she satisfied all the requirements. A student is allowed to enroll in a module of a higher level than her normal level if she meets the requirements. In KSU, a student enrolls in a set of courses online using the institution Web system. Under normal circumstances, all students are registered automatically through the electronic system of KSU following a model plan of study set by the department. This plan includes all prerequisites, maximum and minimum allowable number of credit hours per semester. However, particular students have to ask for help from advisors. The problem encountered by the students is not the enrollment itself, but the decision to be taken before the beginning of each semester concerning how many modules to take and which ones[29].

The research endeavors to exploit KDD tools, especially data mining in order to uncover the most effective sequences of modules that lead to improve student progress and consequently offer students, key elements to make better decision in the enrollment process, specifically those who are stumbling on the road to graduation.

The rationale for the conducted research is presented in the Introduction. The rest of this paper is organized as follows: Section 2 introduces Data Mining usage in the education field. Section 3 presents the basic concepts of decision tree (DT) classification analysis. A review of the related research work is provided in Section 4, the case study, research methodology and the obtained results are described in Section 5. The paper concludes with a summary of the achievements in Section 6.

## 2. DATA MINING IN EDUCATION FIELD

Data Mining (DM) is one of the most prominent topics in educational research fields drawn to explore large amounts of academic data in search of relevant and valuable hidden knowledge [10]. While data mining and knowledge discovery in Databases (KDD) are often treated as synonyms, actually data mining is a stride in the KDD process that involves data analysis and discovery algorithms to generate models upon the data [7]. DM can reveal information that one might never have suspected, useful information that has an impact on decision-making.

Lately, DM methods and tools for analyzing data available at educational institutions, named Educational Data Mining (EDM) [24] [3]; [21] have been widely applied to enhance the goodness of educational system [2] and to solve many problems at higher education such student's retention and dropout, enrollment management, web-based education and student performance [3].

Predicting students' performance using EDM has been of interest for diverse researchers during the last years [3] and it has been demonstrated that DM tools can successfully be utilized to predict student success [10][15]; [15][20]. But many questions are still pending. Indeed, most research on the application of EDM to resolve students' performance problems has been applied primarily to the specific case of students' dropout and grade prediction for a given course based on student demographic data, external assessment or previous courses grades [27]; [25]. Others research have been interested in students' classification based on their profiles [4]; [14]. Yet, no research was done to find most effective sequence of modules for students in higher education. That, what we attempt to uncover in the present work.

Investigations on the important attributes used and methods applied to predict student performance using EDM [27] highlighted that the variables GPA and internal assessment have been the most frequent attributes utilized. And GPA is the main attribute to predict student performance, no doubt because GPA is a measure of student achievement that is internationally recognized and widely used and understood. Moreover, it is demonstrated through the coefficient correlation analysis, to be the most significant input variable [5]. Furthermore, GPA has been revealed as the strongest predictor of student performance [6] 27]. Regarding the methods used, classification has been designed as the most popular task to predict students' performance. Under the classification techniques, Decision Tree and Neural Network [27] are the two methods highly used by the researchers for predicting students' performance.

## 3. DECISION TREE CLASSIFICATION

Classification is a DM technique destined for supervised learning that attributes categories to a collection of data in order to facilitate accurate prediction and analysis. Decision tree (DT) is one of the simplest and yet most successful forms of supervised learning[14] [31];**Error! Reference source not found.** 32] that is successfully used in classification problems [26]. Decision trees are known to be effective methods that achieve good results in practice [27]. It consists of predicting a certain outcome (class attribute) based on a given input. It attempts to discover relationships between the attributes that would make it possible to predict the outcome. Among a suite of algorithms for classification problems in machine learning and data mining, C4.5 is known as the most influential algorithm in DM **Error! Reference source not found.**.The algorithm was proposed in 1992, by Ross Quinlan, to overcome the limitation of the ID3 algorithm (missing values, continuous attribute value ranges, pruning of decision trees, etc.) [22]. C4.5 is an algorithm for inducing classification rules in the form of given trees. It builds DT from a set of training data using the concept of information entropy. In practical DM applications, the input data is expressed as a set of independent instances. These instances are the objects that are to be classified [12]. The instances are the rows of the tables and the attributes are the columns corresponding to data records in database. DT based C4.5 classifies instances by traverse from root node to leaf node. It builds a DM model by creating a series of splits in the tree. These splits are represented as nodes. The algorithm adds a node to the model every time that an input attribute is found to be significantly correlated with the predictable attribute. The way that the algorithm determines a split is different depending on whether it is predicting a continuous column or a discrete column [12][26].

In this study, DT classification based C4.5 algorithm was chosen since DT structure offers the ability to easily generate rules and provide understandable models [7]**Error! Reference source not found.**.

## 4. RELATED WORK

Predicting student's performance using EDM was subject matter of several studies in different universities and countries. For instance, in [19] the authors carried a study to predict the final grade in a specific course using data mining tools. They classified students based on features mined from logged homework data on LON-CAPA a learning online system used in Michigan State University. They conducted two experiments: firstly, they combined different classifiers using the datasets of LON-CAPA in order to obtain significant accuracy and then they used Genetic Algorithm (GA) to optimize performance of this combination of classifiers. In their paper they present and compare the results obtained from the combined classifiers with and without GA. The findings show that using a genetic algorithm leads to minimize the error rate and improves the prediction accuracy at least 10%.

Another similar study **Error! Reference source not found.** was done in CCIS at KSU using data mining techniques to analyze previous students' data in order to predict future students' performance on certain courses. The dataset concerned a sample of Master student records from session 2003 to 2011 in of Computer Science (CS) department. The dataset included student ID, four core courses and 32 elective courses. They used ID3 and J48 decision tree classification methods [22]. They have experiment the performance of each model on elective courses and have compared the two classification models. Results showed a satisfactory performance for each one but the highest accuracy was achieved through J48 classifier. Thus, the latter model was utilized to implement an online system to predict student' grade for a given course. The system recommends also the best courses for students based on two criteria: maximum accuracy value and maximum grade value for all predicted courses.

In [8] the authors explore student demographic data using the KDP (Knowledge Discovery Process) in order to find factors having most impact on student's success. The objective of the study conducted in the University of Mugla Sitki Kocman, Turkey, was to evaluate and develop data-driven method to enhance students' performance in higher education using DM. They developed SKDS (Student knowledge discovery software), an EDM system. They used Decision Tree as classifier to produce rules to feed SKDS. Microsoft Decision Trees (MDTs) in the Microsoft SQL Server Analysis Services were applied to create DT models [16]. The dataset was composed of demographic data of each student and her/his GPA. They classified students in two groups of students' profiles: GPAs ranged from 2.0 to 4.0 and the second group with GPAs of 3.0–4.0. The first and the second classification models show respectively the type of registration to the university and the family income have the most impact on GPA of student. The accuracy of the models was satisfactory with 87 % for Model I and 68 % for Model II.

In [17] Carlos et al. focus on predicting student drop out in compulsory education using DM. They conducted a case study using data from 419 students enrolled in the Academic Unit Preparatoria at the Autonomous University of Zacatecas in Mexico where the dropout rate is the highest of all the educational stages in Mexico. The researchers attempt to detect students' dropout at the most earliest stage of the course. They propose to apply Interpretable Classification Rule Mining (ICRM) as specific classification algorithm at the beginning of a given course and throughout the course progresses. Thus, imbalanced dataset has been gathered at different stage of a given course. They obtained a model for the middle of the course with 80% accuracy and concluded that is a reliable model to classify students at this stage of the course.

Another similar work in [4], the researchers intend to predict student dropout and failure through the whole period of the study in earlier stage using DM and Social Network Analysis (SNA). The study was conducted on bachelor students of Applied Informatics admitted to Faculty of Informatics, Masaryk University, Brno, Czech Republic for the period 2006, 2007, and 2008, three years which correspond to the standard bachelor study. They used two data sets, academic student records enriched by data about their social behavior. The first one contained attributes related to student, semester and other studies. SNA was utilized to compute social behavior data such as intensity of interpersonal communication or number of mutually shared files and neighbors' characteristics. To process the dropout prediction, they applied a certain number of algorithms integrated in the software WEKA. The results show that a highest accuracy is obtained with classifier using both datasets, social behavior data added to academic student data. The authors infer that a student performance seems to be correlated with the social habits, largely with the frequency of communication.

A related work in [29] treats the problem faced by the students to make the right decision in the enrollment process. They proposed a recommender system based on data mining to predict the success or failure of a student in a given course. To feed their system, they examined academic data over seven years of students' enrollment at the School of System Engineering at Universidad de Lima. The data used contained, student information, courses enrolled in, grades obtained, number of courses taken at each academic term, average grade and cumulative grade per academic term. They pre-processed the data and applied the C4.5 classification algorithm. The rules resulting from the classification were utilized by the recommender system to inform the student if her/his enrollment in a given course has good probabilities of success or not. The system does suggestions using collaborative recommendation which consist to compute the similarity of users rather than computing similarity of items [11]. In their case, the system recommends to student to take a particular course based on results obtained by similar profile of students who had taken this course. The dataset used in this research is different from the work set out in the present paper since we consider only courses enrolled in each term and the average grade at the beginning and the end of each academic term. Our objective is to uncover patterns related to most effective sequences of modules in student's academic itinerary.

Another related work in [6] using MATLAB 2015b platform [18], the authors perform prediction and classification from graduated student data recorded between 2007 and 2011 in the Computer Science (CS) department at Faculty of Science and Technology, Sakon Nakhon Rajabhat University, Thailand. Two models for the 3rd year and 4th year have been constructed to predict grades results by applying Neural Network algorithm [26] and seven classes of students who had similar grade pattern in each course have been obtained using the clustering technique namely K-means algorithm [12]. The outcomes show that clusters of students might serve to find a possible study path for remaining semesters.

# 5. CASE STUDY

This study was initiated because of the ongoing challenges facing students to choose the best sequence of courses to take within the semester according to their preferences and prerequisites courses. It aims to find which sequences of modules are associated with academic success using DM techniques. The steps of the study are described below.

## 5.1 Context and Problematic

The study concerns students of Bachelor and was carried out in the department of (IT) at the College of Computer and Information Science (CCIS) at (KSU). The program is an on-campus, daytime, and in-class program. The Department follows the semester system. Two semesters are offered in each academic year. Each semester is also called level. The first year (2 levels) corresponding is not considered in this study since it is a Preparatory Year (PY). After completing the PY, students are admitted to the college and distributed to the various departments according to three criteria: their preference, Grade Point Average (GPA) from the PY, and the capacity of each department. Students have the opportunity to start focusing on a specific concentration of their choice after their 5th semester. The (IT) department offers three concentrations (tracks), thereby allowing students to gain their degree of Bachelor of Science in Information Technology in any of the following concentrations: Data Management (DAM), Web Technology and Multimedia (WTM) and Networks and Security (NS).

The (IT) Department where the study was done offers several courses (e.g. Software Engineering, Database, Intelligent Systems, etc.). The students have to choose among core and elective courses, a total number of 51 core and more than 25 elective courses. The credit hours for each student during a semester must be more than eleven hours and less than twenty hours. All students at IT department are required to maintain a GPA of at least 2.75 out of 5.0. Every IT student needs to choose at least one of the three above majors. Before the beginning of the semester, students should prepare their study schedules carefully to meet the department, college and university requirements. An electronic Plan of Study is automatically created by Edugate, the electronic system of the KSU based on the student's program curriculum. The registration system tracks the courses that have been completed by each student, semester-by-semester GPA, as well as the overall accumulative GPA, in addition to other academic functionalities. Under normal circumstances, all students are registered automatically through Edugate following a model plan of study set by the department. This plan includes all prerequisites, maximum and minimum allowable number of credit hours per semester. The system allows the student to make changes and adjustments within the preset rules [28]. Still, some undergraduate students have lack of success at mid-term stage. Some students are unsuccessful because they miss making the right choice of courses

## 5.2 Research Methodology

Advice and guidance to students in higher education should be improved to facilitate modules' sequencing selection. For example, students need to know that some combinations of courses are not good or not suitable for them. Our methodology tries to find which sequences of modules are most effective for students at each level of their academic itinerary. In the present study DM methods are applied to graduated students' data with the goals to answer the following two research questions:

- What are the most effective sequences of modules for students to improve their performance? In other words can we find dependency between sets of courses taken by student in a given semester with her performance?
- Can we classify the sequences of courses that have been taken by students? If so, can we show how bad sequences impact student's performance. Can we help institutions and advisers to guide students more efficiently and effectively?

In this paper, regarding these questions, we intend to find similar patterns of use in the data gathered from the database of the institution, and eventually be able to make predictions as to the most-propitious path of studies for student based on their present usage. Next, the different steps of our methodology are detailed.

## 5.3 Data Collection

This case study focuses on female students studied in the regular bachelor track who are graduated from IT department. Our dataset is taken from the system Edugate. The data set contains information of **647** students; each student data includes at least eight semesters of study. Each semester data includes date, student's ID, student's national ID, semester's date, student's major, semester and cumulative scores and at least 4 tuples. Each tuple includes module's name and code taken by the student along with their grading and GPA before the start and the end of each level. Our study was conducted based on dataset analysis from 2007 to 2013, which is approximately 12 semesters.

## 5.4 Data Preparation and Preprocessing

The first stage made in the tremendous research field of DM consists in an initial data exploration. For building various models and choosing the best ones, based on their predictive performance, it is necessary to perform a preliminary exploration of the data to better understand its characteristics.

**Table 1. Features selection**

| Feature | Description |
|---|---|
| ID | Unique Student's identification |
| GPA_ Before | Grade Point Averages of student before taking the courses of the semester to be enrolled in. |
| Level | Level of student when taking this course. |
| Track | Track followed per student. |
| C_code | Course code |
| C_name | Course name |
| C_Type | Core course or elective |
| C_Type_Track | Course core in the track or optional. |
| Grade | Student's grade for this course |
| GPA_After | Grade Point Averages of student at the end of the semester |
| CGPA | Cumulative Grade Point Average |

Preparing input for a DM investigation generally consumes the most of the effort invested in the entire DM process [12]. As a first step towards creating a dataset for the study, an export of the database created an excel file, including the information shown in table 1.

In the second step, some of the attributes have been removed, e.g., student's national ID, fields containing data that is of no interest to the research. From the data collected, tuples related to students having more than eight levels were deleted. Only regular curriculum was considered. GPA before and GPA after were transformed into one attribute called GPA.

GPA = GPA_After – GPA_Before. If the result is positive (student progress positively) the label is **P** and **N** if it is negative. The attributes C_name, C_Type, C_Type_Track and Grade were discarded. At this stage, features selection is set to student id, course code, level, track and GPA. For each, student, and each track and each level, there are a number of tuples. And each tuple corresponds to the courses taken by the student at this level, in this track. We recall, in this study, we investigate which sequences of courses are better for each level and each track. The tuples of courses for each level and track for a student were transformed in a sequence of courses by merging them. For example, there is a student, with track WTM and has eight levels. In each level, there are four courses, each one described by one raw, which set 24 tuples for this student. After merging rows of the same level, there will be only 6 tuples for this student. By doing this transformation, the data is reduced and the sequence of courses is much more apparent. This was done by VBA macros using Excel functions. Each unique sequence of courses is uniquely named. For example, **S4-1** represents the sequence of courses for level 4 and numbered 1, **S4-2** represents another sequence of courses for level 4 and numbered 2 and so on. For each level, a number of sequences of courses were found as shown in an example in table 2.

**Table 2. Features selection**

| **Sequence of courses' codes concatenated** | **Attrib**. |
|---|---|
| IC102CSC113IT221IT222MATH244STAT324 | S4-1 |
| CSC113IT221IT222MATH244STAT324 | S4-2 |
| IC103CSC113IT221IT222MATH244STAT324 | S4-3 |
| IC102CSC113IT221IT222STAT324 | S4-4 |
| IC102CSC113IT221MATH244STAT324 | S4-5 |
| IC102IT222STAT324PHY104 | S4-6 |
| IC102IC103CSC113IT221IT222MATH244STAT324 | S4-7 |
| IC102CSC111MATH244STAT324ARAB103 | S4-8 |
| IC102CSC113IT221IT222 | S4-9 |
| IC102CSC113IT221IT222MATH244 | S4-10 |
| IC102CSC113IT221IT222MATH244PHY104 | S4-11 |

Mainly, the challenge in the presented study is to predict the student performance based on the collection of attributes providing information about the sequence of modules taken by the student at each level and each track. The concept to be learned by DM algorithm is the "GPA", positive or negative thus it has been selected as the target attribute in this case.

First, an attempt to create a classifier with a common dataset for all levels and track was made but the result was mediocre with very low accuracy and understandable DT. Hence, for precision and accuracy concerns, we search for sequences of courses in each semester and track separately. Thus, data was divided into 3 groups corresponding to common levels: 3, 4, 5 and 9 groups corresponding to levels 6, 7 and 8 including tracks NSN, WTM, DAM for each semester. 12 excel files were saved, as Comma Delimited (.csv) files and then transformed to an Attribute Relation File Format (.arff) file which is better when extracting the knowledge using DM techniques. The final dataset used for the current study all levels and track combined contains 4800 instances.

After data cleaning step and transformation of data into a more workable format, the next step is to conduct tests in order to: firstly construct models of better sequences of modules, secondly map and explore dependency of modules students are enrolled in and their performance which is presented by the GPA attribute since the success of each student is reflected by his/her GPA and last validate findings.

## 5.5 Data Modeling

During this phase, the methods for building a model that would classify the sequences of modules taken by students into the two classes (P/N) based on previous students' academic records, are considered and selected. The reached research results are presented in the next section.

## 5.6 J48 classification

In our experimental study, the WEKA [30] J48 classification filter is applied. It is an implementation of C4.5 decision tree algorithm which is suitable for our type of data. In addition, it is a highly ranked algorithm in DM research according to **Error! Reference source not found.** and has shown in many studies its potential to yield good results [12]. Experiments have been carried out for each level and each track.

The goal is to find a model able to predict the class: GPA in a correct manner. To attain this objective, a decision tree was built. Performance was measured in terms of accuracy (the number of correctly classified examples over the number of all examples) and True Positive Rate (the proportion of examples which were classified as class x, among all examples which truly have class x). Both 10-fold cross-validation and percentage split testing have been used. The achieved results are slightly better for the percentage split testing option but not significant difference.

## 5.7 Classification results

By applying J48 classifier to the different datasets corresponding to levels and tracks separately, models have been obtained of almost 65% accuracy which is quite satisfying in our field of research. These models could be applied to predict the academic performance of the following generation of students. The J48 classifier classifies correctly at least 2/3 of the instances (65.58 % for the 10-fold cross-validation testing and 66.59 % for the percentage split testing) for each dataset.

The overall results of classifiers' performance on our datasets are shown in the Table 3 and 4.

**Table 3. Accuracy percentage for levels: 3, 4 and 5**

| **Level** | **Accuracy** | **TP** |
|---|---|---|
| 3 | 70.6 | 66.53 |
| 4 | 80.76 | 80.8 |
| 5 | 85.5 | 79.7 |

**Table 4. Accuracy/TP percentage for each level of each track**

| **Track**<br>**Level** | **WTM** | **DAM** |
|---|---|---|
| | **Accur. / TP** | **Accur. / TP** |
| 6 | 65.58 / 63.95 | 85.5 / 80.53 |
| 7 | 70.66 / 68.03 | 88.66 / 87.3 |
| 8 | 87.52 / 87.0 | 90.98 / 90.0 |

The tables 3 and 4 show the performance of the algorithm J48 is satisfactory with accuracy above 65% and TP above 60%.

## 5.8 Discovered models

The present exploratory analysis of the 2007-2013 dataset reveals many models useful for recommendations and guidance. Different models at each level and for each track discovered by our investigation are shown in table 5, 6, 7 and 8.

DM enabled us to accurately identify most effective models of sequences of courses to be recommended to students. The generated knowledge will be quite useful for understanding the problem of sequencing in a better way. The advisor would be able to guide students in a timely manner.

**Table 5: Most effective sequences leading to students' progress for level 3, 4 and 5**

| | |
|---|---|
| L3 | IC101, CSC111, MATH106, MATH151 |
| L4 | IC102, CSC113, IT221, IT222 |
| | IC102, CSC113, IT222, MATH244, STAT324 |
| | IC103, CSC113, IT222, MATH244 |
| L5 | IC107, IT211, IT311, IT322 |

**Table 6: Most effective sequences leading to students' progress for levels 6, 7 and 8 for track WTM**

| | |
|---|---|
| 6 | IT321, IT323, IT324, IT325, IT341 |
| 7 | IT419, IT422, IT496, IT342, IT443<br>IT419, IT422, IT496, IT342, IT454<br>IT419, IT422, IT496, IT342, IT452 |
| 8 | IC108, IT497, IT424, IT361, IT499<br>IC108, IT497, IT424, IT443, IT499<br>IC108, IT497, IT444, IT499<br>IT497, IT352, IT999<br>IT497, IT455, IT999<br>IT497, IT361, IT499 |

**Table 7: Most effective sequences leading to students' progress for levels 6, 7 and 8 for track: DAM**

| | |
|---|---|
| 6 | IT321, IT323, IT324, IT325, IT331 |
| 7 | IT419, IT422, IT496, IT434, IT361<br>IT419, IT422, IT496, IT434, IT332<br>IT419, IT422,IT424, IT434, IT361 |
| 8 | IC108, IT497, IT424, IT444, IT499<br>IC108, IT496, IT497, IT499<br>IC108, IT496, IT497, IT999<br>IT497, IT352, IT999<br>IT497, IT455, IT999<br>IT497, IT361, IT499 |

**Table 8: Most effective sequences leading to students' progress for levels 6, 7 and 8 for track: Network**

| | |
|---|---|
| 6 | IT321, IT323, IT324, IT325, IT351<br>IT321, IT323, IT325, IT351 |
| 7 | IT419, IT422, IT496, IT453<br>IT419, IT422, IT496, IT424 |
| 8 | IC108, IT496, IT497, IT499<br>IC108, IT496, IT497, IT999<br>IT497, IT424, IT325, IT999<br>IT497, IT424, IT321, IT999<br>IT497, IT352, IT999<br>IT497, IT455, IT999<br>IT497, IT361, IT499 |

Regarding certain courses not appearing in the most effective sequence of modules, the institution will be able to improve the program. For example, the module IT 211 (Assembly) can be more beneficial for learning if combined with IT 321 (Computer Architecture) and similarly IT224 (Network 1) can be combined with IT424 (Network 2) to yield better results. ARAB101 can be moved to PY. We note that the grades of IT211 and IT224 are bad in most semesters of the study and consequently impacted negatively the GPA of the semesters where they have been taken by students.

## 6. CONCLUSION

This research has sought to find the most effective sequences of modules that lead to students' performance progress based on real-world graduated students' data in curriculum Information Technology. The proposed methodology has shown to be valid for predicting sequences of courses leading to student performance progress in higher education. We carried out experiments using data from 647 graduated Bachelor students.

It worth to notice that most of the previous and current research on the application of EDM to resolve the students' performance problems has been applied primarily to the specific case of students 'dropout or predicting students' performance based on student demographic and external assessment. However, no research was done to find most effective sequences of modules in higher education.

This work discovers classification models trustworthy enough to make an early prediction of effective sequences of modules to be taken by students before the beginning of the semester. In fact, we obtained good results.

Furthermore, this study has offered us an opportunity to reflect as academic researcher that Data mining is an efficient analytical instrument that permits educational institutions to better, proactively manage student outcomes to overcome the problems that led to falling and to follow-up educational achievement of students stumbling.

Therefore, the obtained models can be used as guidance to students by academic advisors. As future work, a recommender system based on data of previous sequences of modules taken by students will be developed to suggest for each student a suitable pathway of her/his academic study. Of course, further tests and experiments with new records have to be done to ensure the efficiency of these models.

## 7. ACKNOWLEDGMENTS

This research project was supported by a grant from the “Research Center of the Female Scientific and Medical Colleges”, Deanship of Scientific Research, King Saud University.

## 8. REFERENCES

[1] Al-Saleem M., Al-Kathiry N., Al-Osimi S. and Badr G. “Mining Educational Data to Predict Students' Academic Performance”. Springer, 2015.

[2] Bousbia, N., Belamri, I: Which Contribution Does EDM Provide to Computer-Based Learning Environments?. Educational Data Mining: Applications and Trends. Springer, 2014.

[3] Baker, R., Yacef, K. “The State of Educational Data Mining in 2009: A Review and Future Visions”. Journal of Educational Data Mining, Vol. **1**, Issue 1, pp.3-17, October 2009.

[4] Bayer J. Bydzovska, H., Geryk, J., Obsivac, T., Popelinsky, L. "Predicting dropout from social behavior of students". Proceedings of the 5th International Conference on Educational Data Mining, Crete, Greece, pp.103–109, 2012.

[5] Bin Mat, U., Buniyamin N., Arsad P. M., Kassim R., An overview of using academic analytics to predict and improve students’ achievement: A proposed proactive intelligent intervention, in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, 2013, pp. 126–130.

[6] Chanamarn, N., Tamee, K. "Enhancing Efficient Study Plan for Student with Machine Learning Techniques", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.3, pp.1-9, 2017.DOI: 10.5815/ijmecs.2017.03.01

[7] Fayyad, U. Piatetsky-Shapiro, G. and Smyth, P. “From Data Mining to Knowledge Discovery in Databases”. AI Magazine Volume 17 Number 3, 1996.

[8] Guruler, H. and Istanbullu, A. “Modeling Student Performance in Higher Education Using Data Mining”. A. Peña-Ayala (ed.), Educational Data Mining, Studies in Computational Intelligence 524, DOI: 10.1007/978-3-319-02738-8_4, Springer International Publishing Switzerland 2014

[9] Grumbach, S. Valduriez, P. "Les données en question". [Online]. Available: https://interstices.info/jcms/p_84069/les-donnees-en-question [Accessed on 2 Oct. 2017].

[10] Hall, M., Eibe F., Holmes, G, Reutemann, Bernhard P. and Witten, Ian H. Data mining with WEKA, update.

[11] Hssina, B., Merbouha, A., Ezzikouri, H. and Erritali, M. A comparative study of decision tree ID3 and C4.5. International Journal of Advanced Computer Science and Applications, 4(2), 13-19, 2014.

[12] Herlocker J.L., Konstan J.A., Terveen L.G., Riedl J.T. “Evaluating Collaborative Filtering Recommender Systems”, ACM Transactions on Information Systems, Vol. 22, No. 1, January 2004, Pages 5–53.

[13] Ian.H. Witten, E. Frank, Mark A. Hall. “Data Mining: Practical Machine Learning Tools and Techniques”. 3rd edition, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©2011.

[14] Ian.H. Witten, E. Frank, Mark A. Hall. “Data Mining: Practical Machine Learning Tools and Techniques”. 3rd edition, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©2011.

[15] Josip Mesarić, and Dario Šebal, Decision trees for predicting the academic success of students, 2016.

[16] Kabakchieva D. “Predicting Student Performance by Using Data Mining Methods for Classification”. Cybernetics and Information Technologies • Volume 13, No 1, 2013.

[17] Larson, B., English, D., Purington, P. “Delivering Business Intelligence with Microsoft SQL Server 2012”. McGraw-Hill, New York, 2012.

[18] Márquez-vera, C., Cano, A., Romero, C , Noaman, A., Fardoun, Habib M.and Ventura, S. “Early dropout prediction using data mining: a case study with high school students”. Expert Systems, Vol. 33, No. 1, February 2016.

[19] MATLAB, “MATLAB Environment”, from http://www.mathworks.com/products/matlab/, 2016.

[20] Minaei-bidgoli, B., Kashy, D. A., Kortemeye,r G., Punch, W.F. “Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA”. 33rd ASEE/ IEEE Frontiers in Education Conference, Nov. 5-8, 2003, boulder, co.

[21] Mueen, A. , Zafar, B., Manzoor, U. “Modeling and Predicting Students' Academic Performance Using Data Mining Techniques”. I.J. Modern Education and Computer Science, 2016, 11, pp. 36-42 Published Online November 2016 in MECS, DOI: 10.5815/ijmecs.2016.11.05

[22] Pena-Ayala, A. “Educational data mining: applications and trends”. 2013

[23] Quinlan, R. J. (1996). Improved Use of Continuous Attributes in C4.5. Journal of Arti, 4, 77-90.

[24] Romero, C., S. Ventura. Educational Data Mining: A Survey from 1995 to 2005. – Expert Systems with Applications, Vol. 33, pp.135-146, 2007.

[25] Romero, C., Ventura, S. Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 40(6), pp. 601–618, 2010.

[26] Romero, C., & Ventura, S. “Data mining in Education,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), pp.12–27, 2013.

[27] Russel S., Norvig P. “Artificial Intelligence: A novel Approach”. Third edition, 2010.

[28] Shahiri A., Husaina W., Abdul Rashida,N.. “A Review on Predicting Student’s Performance using Data Mining Techniques”. Procedia Computer Science 72, pp. 414 – 422, 2015.

[29] SELF-STUDY REPORT Bachelor of Science Program in Information Technology July 2012.

[30] Vialardi C., Bravo J., Shafti, L., Ortigosa, A. “Recommendation in Higher Education Using Data Mining Techniques”. International Conference on Educational Data Mining (EDM) (2nd, Cordoba, Spain, Jul 1-3, 2009).

[31] Weka, 2017. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/.

[32] Xie N., Liu Y. “Review of decision trees”, in: Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 5, 2010, pp. 105–109.

[33] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S.Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, 2008. "Top 10 algorithms in data mining", Knowledge and Information Systems, 14, 1: 1–37.