Research Article

International Journal of Advanced Computer Research, Vol 8(36)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2018.836003

A proposed academic advisor model based on data mining classification techniques

Mohamed Hegazy Mohamed[1*] and Hoda Mohamed Waguih[2]
Demonstrator, October 6 University, Giza, Egypt[1]
Associate Professor, Department of Computer and Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt[2]

Received: 20-December-2017; Revised: 29-January-2018; Accepted: 27-April-2018

Abstract
University and higher institute admission are an intricate decision process and it is an important responsibility of the students to select the correct study track. The increase of the student's major dropout rate in higher education systems is one of the important problems in most institutions. One approach to solve such problem and succeed in academic life is to help the students in selecting a suitable major and assign them to the right track. The objective of our research is to build academic advisor model to students for their higher education which utilize classification data mining for recommending the suitable academic major. The method applied in the research is data mining classification techniques through decision tree method for advising students to select suitable major and help assign them to the right track. The proposed model classifies students and matches them to the proper study tracks according to their features. The three decision tree classification algorithms, namely J48, random tree and reduces error pruning (REP) tree was first applied to real data in a managerial higher institute in Giza Egypt and results are compared between them. Finally, the results showed that J48 algorithm gives 16 rules and we eliminate the rules that give low CGPA and we will use the 5 better rules that have the highest CGPA based on CGPA grade that equal (A) and J48 algorithm gives the highest accuracy 87.64% and classification error was 12.36% and was thus selected as the main classifier for building the proposed model based on the rules that we obtained from J48 algorithm than the two other classification algorithms and thus suggest using the generated J48 decision tree in our proposed student advising model to enhance students' academic performance and decrease dropout.

Keywords
Data mining, Classification, Decision trees, Higher education, J48, Random tree, REP tree.

1.Introduction
The growth of academic data in higher education systems increases rapidly. This data is a strategic resource for higher education institution. Making the most use of these strategic resources will lead to improving students' performance, thereby, improve quality of whole educational processes. Analysis of this data needs powerful tools such as data mining to extract the meaningful information from large data using some algorithms [1]. A major problem that faces higher educational institutes is the increase in student dropout. To solve such problem and succeed in academic life can be done by helping students selecting a suitable major and assign them to the right track using data mining classification techniques.

In the present study, we investigate how to use the data mining classification technique in advising students to select suitable major and help assign them to the right track in the first academic year. As part of our investigation, we conducted an experiment using three classification algorithms, namely J48, random tree and REP on real data in a managerial higher institute in Giza Egypt.

In the next section, we discuss some of the literature review addressing the problem of students dropout in higher education. Section 2 discusses materials and methods which explore the data mining classification technique. Results are introduced in section 3. Discussion and findings are addressed in section 4. Finally, conclusions and future recommendations were discussed in section 5.

This section will discuss some of previous research that related to this problem and how they applied data

*Author for correspondence

mining classification techniques for enhancing students' academic performance and decreasing dropout of students in educational systems.

In [2], the authors use three algorithms; J48, Naive Bayes, and Random Forest to predict the success rate of students enrolled in a course using machine learning classification algorithms. The main objective of their methodology was to examine the major factors which are causing a dropout of students and to help students improve their performance by evaluating themselves on the basis of their prior records. The results of the study show that their proposed model was helpful in predicting the students' result on behalf of their performance in prior tests. The algorithm J48 shows that 93.3% instances were correctly classified while Naive Bayes presents 86.6% accuracy and Random Forest depicts that 100% instances were correctly classified.

In [3], the authors built a classification model to predict students' performance in higher education institute. They used the NBTree data mining classification technique and conducted several experiments to discover a prediction model for students' performance. The class labels of students' performance were students' status at study, graduate's predicates, and length of study. The experiments were conducted with the two-level classification, the university level and faculty level. Their resulted model indicated that some attributes had significant influence over students' performance.

In [4], the authors proposed a recommender system which provides assistance to students for selecting their study stream in higher education. The system uses classification and clustering techniques in recommending the right academic stream and colleges to students. In their system, they group the students into a number of clusters and match their profile with the more relevant cluster using FCM algorithm and C4.5 classification algorithm for classifying the students into the course and the college which matches to them. The advantage of their system was the accuracy of the prediction and the speed of the results provided. The data mining techniques used and the performance analyzed in their study was using the WEKA tool.

In [5], the authors describe their initial efforts to model student dropout using a large dataset on higher education attrition, which tracks over 32,500 students' demographics and transcript records at one of the nation's largest public universities. The results of their study highlight several early indicators of student attrition and show that dropout can be accurately predicted even when predictions are based on a single term of academic transcript data.

In [6], the authors applied different data mining approaches for the purpose of examining and predicting student dropouts through their university programs. In their study, they use select a total of 1290 records of computer science students graduated from ALAQSA university between 2005 and 2011. In order to classify and predict dropout students, they use train different classifiers on the used datasets including decision tree (DT), naive Bayes (NB) and tested using 10-fold cross validation. The accuracy of the DT, and NB classifiers were 98.14% and 96.86% respectively. Their study also includes discovering hidden relationships between student dropout status and enrollment persistence by mining a frequent case using the FP-growth algorithm. Their study finds that mastering "digital design" and "algorithm analysis" courses has a great effect on predicting student persistence in the major and decrease student's likelihood of dropout.

In line with the previously mentioned studies, the main objective of the present study focused on how to use data mining classification techniques for predicting student dropout. In this paper, we used three decision tree algorithms on the collected real dataset; conduct a comparative study on the obtained results so as to choose the most accurate algorithm in predicting the student major based on the selected features, then, used of J48 algorithm and filtering the most important rules.

2.Materials and methods

Various algorithms and techniques like classification, clustering, regression, artificial intelligence, neural networks, association rules, decision trees, genetic algorithm, nearest neighbor method etc., were used for knowledge discovery from databases. Classification is one of the most frequently studied problems with data mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). Classification is also a data mining technique that maps data into predefined groups or classes. It is a supervised learning method which requires labelled training data to generate rules for classifying test data into predetermined groups or classes. It is a two-phase process. The first phase is the learning phase, where the training data are analysed and

classification rules are generated. The next phase is the classification, where test data are classified into classes according to the generated rules. Since classification algorithms require that classes be defined based on data attribute values, we created an attribute "class" for every student and there are different classification methods like decision tree. The decision tree structures are a common way to organize classification schemes. In classifying tasks, decision trees visualize what steps are taken to arrive at a classification. Decision trees are the classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data [7].

J48 algorithm is a successor to iterative dichotomiser (ID3) developed by Quinlan Ross. It is also based on Hunt's algorithm and handles both categorical and continuous attributes to build a decision tree. In order to handle categorical attributes, it splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses gain ratio as an attribute selection measure to build a decision tree. It removes the biases of information gain when there are many outcome values of an attribute. Also, J48 calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximized. It uses a pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification. Entropy and information gain measures are used by J48 to construct a decision tree. J48 algorithm flowchart has been shown in *Figure 1*.

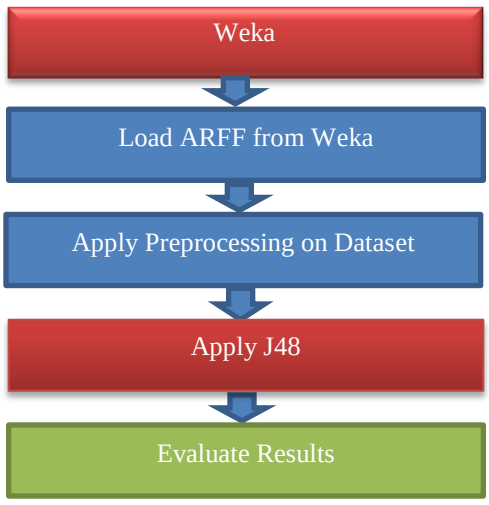

Figure 1 J48 Algorithm flowchart

This operator uses only a random subset of attributes for each split. This algorithm works exactly like the decision tree algorithm with one exception: for each split only, a random subset of attributes is available. It is recommended to study the documentation of the decision tree algorithm for basic understanding of the decision trees. The random tree algorithm learns decision trees from both nominal and numerical data. Decision trees are powerful classification methods which can be easily understood. The random tree operator works similar to Quinlan's C4.5 or CART but it selects a random subset of attributes before it is applied. The size of the subset is specified by the subset ratio parameter. Representation of the data as tree has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of the label based on several input attributes of the example set. Each interior node of the tree corresponds to one of the input attributes. The number of edges of an interior node is equal to the number of possible values of the corresponding input attribute. Each leaf node represents a value of the label given the values of the input attributes represented by the path from the root to the leaf [8]

The REP tree algorithm is the fastest decision tree learner. It Builds a decision or a regression tree using information gain or variance and prunes it using reduced-error pruning (with back fitting), and only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5). Random tree uses class for constructing a tree that considers K randomly. We have a random subset of attributes to deal with the limitation of the decision tree. The value of random subset is based on operator, in this way we can solve the classification as well as regression and prediction problem.

In this research we conducted an experiment in a higher education system where three decision tree algorithms, namely J48, random tree and REP tree were used in real instance examples representing the student's records of the first academic year from the managerial higher institute 'Tammoh' in Giza, Egypt. A total of 8080 records and 7 attributes throughout the years 2007 to 2016 are selected from the student's database. Student data in the first academic year included personal information and student academic qualifications. The attributes describing the selected dataset are student's genders, place of birth, gender, high school major, and high school grade, university major and cumulative grade point average (CGPA)

which was selected in our experiment to represent the decision attribute. The detailed description of the attributes is presented in *Table 1*.

Data preparation and pre-processing are very important steps in any data mining process that usually consumes the bulk of the effort invested in the entire data mining process. Some pre-processing was conducted on the selected dataset prior to the mining process. In this context, we ignored some tuples to handle missing values in some of the conditional attributes and eliminated some irrelevant attributes which have a minor effect on the generated classifiers or may result in over fitting problem. Data tuples from multiple sources were then merged into a coherent data source and a discretization process was applied to transform some attributes such as age, high school grade and CGPA from numeric to nominal attributes as shown in *Table 1*. Other pre-processing was made on some of the attributes in the dataset after consulting some experts in higher education. Some of them were narrowing the HS_Major attribute domain from 6 to 3 values by merging the management and services, commercial, industria_3_years, industria_5_years into one domain namely Other. Also, we narrowed CGPA attribute domain from 9 to 4 by merging A+, A and B+ to Excellent value, B and C+ to Good, and the values C, D+ and D to Fair. Finally, prior to applying the selected classification algorithms, the target dataset was transformed to the specific input data formats used by the selected data mining tool.

Table 1 Student attributes

ID	Attribute	Description		Domain
1	Gender	Student's gender Binary: 'F' - female or 'M' - male)		M F
2	Place of birth	Student's place of birth Nominal: 'U' = urban or 'R' = rural		U R
3	Age	Student's age Nominal [A:18-20 B:21-22 C:23-25]		A B C
4	HS_Major	High School Major: Nominal		L S O
		Literary_General_Secondary = L Scientific_General_Secondary = S		
		Management and Services = 'MS' Commercial = 'Com' Industria_3_Years = 'I3' Industria_5_Years = 'I5'	Other = O	
5	HS_Grade	High School Grade: Type Nominal A = Excellent from 85%: 100% B = Good from 61%: 84% C = Fair from 50%: 60%		A B C
6	CGPA cumulative grade point average	High Excellent >=93 &<100 A+ Excellent >=85 &<93 >=80 &<85 A High Very Good >=80 &<85 B+	A = Excellent	A B C
		Very Good >=75 &<80 B High Good >=70 &<75 C+	B = Good	
		Good >=65 &<70 C High Acceptable >=60 &<65 D+ Acceptable >=50 &<60 D	C = Fair	
		Failure >=0 &<50 F	F = Fail	F
7	Major	Student's Major (Accounting = ACC Management = Manag Management Information System = MIS)		ACC MIS Manag

The three decision tree algorithms used in our experiment were implemented using WEKA; an open source data mining software that contains Java implementations of many popular machine learning-algorithms including some popular classification algorithms. To use WEKA, the collected data need to

be prepared and converted to (Arff) file format to be compatible with the WEKA data mining toolkit [8]. It uses a wide variety of descriptive and predictive techniques to give you the insight to make profitable decisions. During the modelling phase, modelling techniques was selected and applied to the dataset used in the study. This phase included selecting appropriate modelling technique, building the models and final assessment of the model. Subsequently, the model selection involves selecting appropriate techniques for the problem; refine the models whenever is necessary in order to meet the requirements.

In this work we build an architecture model for an academic advisor to guide students in selecting his/her suitable major and assign them to the right track in higher education system. The model is composed of two main phases as shown in *Figure 2*. In the first phase, the three algorithms are applied to the target dataset and the generated decision trees are compared to select the best classifier as discussed earlier. In phase 2, the selected classifier is used to guide the student in selecting the suitable major that produce the highest CGPA which is our main objective.

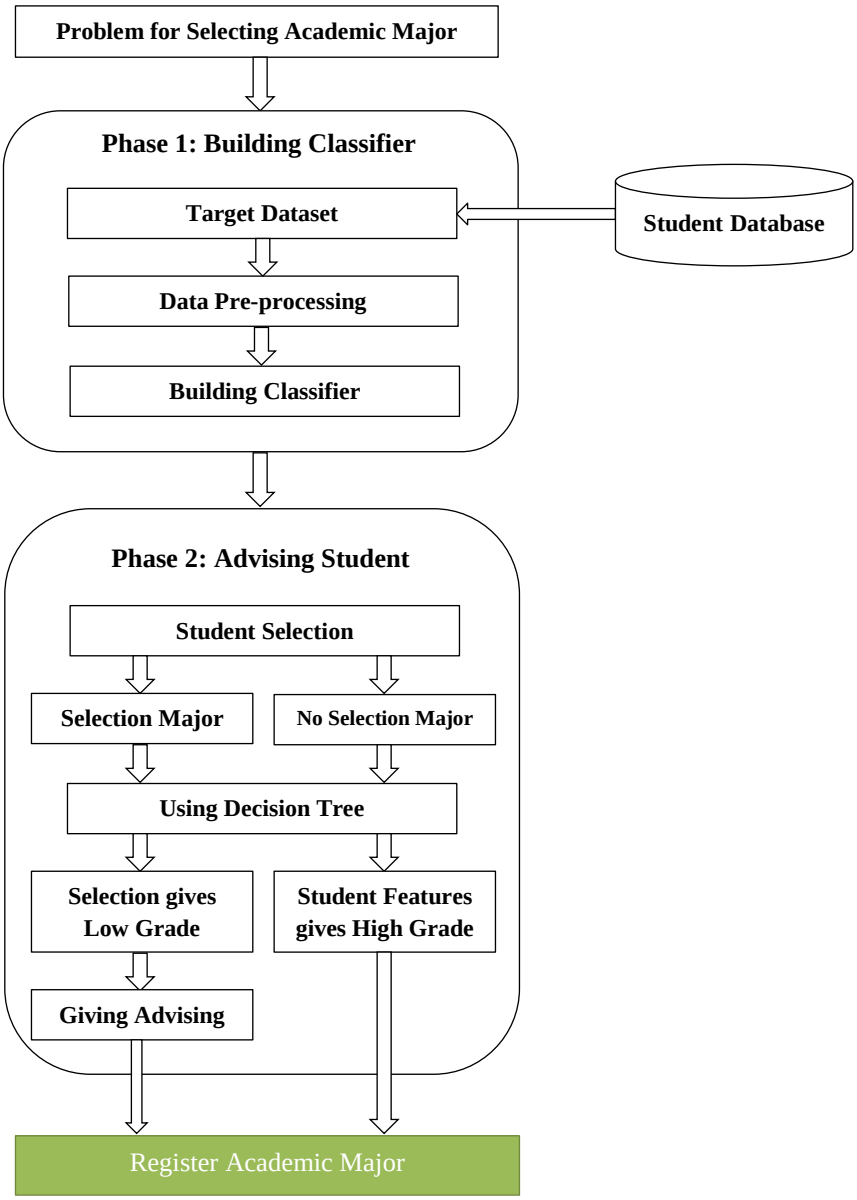

Figure 2 Architecture model for student academic advisor

In order to achieve this objective, we must consult the classifier in one of two scenarios. In the first scenario, a student selects a major. If the selected major, according to the student attributes, produces low CGPA, we advise him to change his selection. Otherwise, if student attributes provide high CGPA, we register his selected major directly. The second scenario, the student has no selection and request assistance in selecting a suitable major. In this case, we consult the classifier and give him an advice that results in higher CGPA according to his attributes.

3.Results

In our experiment we used CGPA as a decision attribute to generate the classifiers based on the selected decision algorithms. The reason for choosing the CGPA is to advise students to select the suitable major that will result in the highest possible CGPA. The three generated classifiers are then compared and the best one in term of accuracy of results are chosen to be used in the proposed model. In this context, J48 algorithm outperforms the two other algorithms; random tree and REP tree, as it provides the highest accuracy 87.64 % with a classification error of 12.36 % as shown in *Table 2* and provides the best decision tree as shown in *Figure 3*. The generated decision tree as showed in *Figure 3* showed that the attribute major was selected as the root node, and we obtained 16 rules as showed in *Table 3* that help students to choose the right track. Finally, we obtained the best 5 rules as showed in *Table 4*. It helps us to give student's advising early. The results of applying the three algorithms were analysed and compared to select the best generated decision tree in terms of accuracy of the classifier. The results showed that J48 algorithm gives the highest accuracy 87.64% and classification error was 12.36% than the random tree and the REP tree algorithms and was thus made the J48 algorithm as the main classifier for building the proposed model.

Table 2 Classification results for three algorithms

Algorithm	Time /Sec	Model evaluation			
		Correctly classified		Incorrectly classified	
		#	%	#	%
J48	0.05	7081	87.64	999	12.36
Random tree	0.02	7065	87.43	1015	12.56
REP tree	0.03	7065	87. 44	1015	12.56

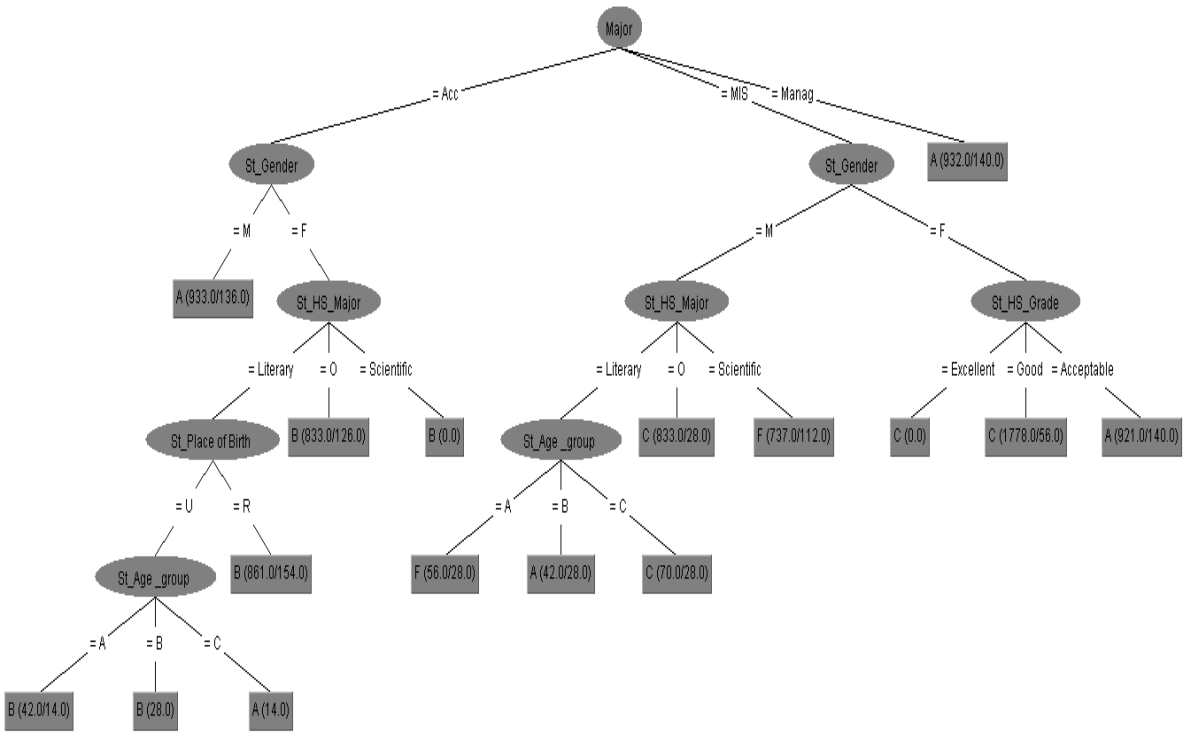

Figure 3 J48 Decision tree

Table 3 J48 Classifier rules output J48 pruned tree

```
MAJOR = ACC
|  ST_GENDER = M: A (933.0/136.0)
|  ST_GENDER = F
|  |  ST_HS_MAJOR = LITERARY
|  |  |  ST_PLACE OF BIRTH = U
|  |  |  |  ST_AGE _GROUP = A: B
|  |  |  |  ST_AGE _GROUP = B: B
|  |  |  |  ST_AGE _GROUP = C: A
|  |  |  ST_PLACE OF BIRTH = R: B
|  |  ST_HS_MAJOR = O: B
|  |  ST_HS_MAJOR = SCIENTIFIC: B
MAJOR = MIS
|  ST_GENDER = M
|  |  ST_HS_MAJOR = LITERARY
|  |  |  ST_AGE _GROUP = A: F
|  |  |  ST_AGE _GROUP = B: A
|  |  |  ST_AGE _GROUP = C: C
|  |  ST_HS_MAJOR = O: C
|  |  ST_HS_MAJOR = SCIENTIFIC: F
|  ST_GENDER = F
|  |  ST_HS_GRADE = EXCELLENT: C
|  |  ST_HS_GRADE = GOOD: C
|  | ST_HS_GRADE = ACCEPTABLE: A
MAJOR = MANAG: A
```

Table 4 J48 Classifier rules output J48 pruned tree

ID	Rules
1	If Major = Acc and St_Gender = M then CGPA = A
2	If Major = Acc and St_Gender = F and St_HS_Major = Literary and St_Place of Birth = U and St_Age _group = C then CGPA = A
3	If Major = Acc and St_Gender = F and St_HS_Grade = Acceptable then CGPA = A
4	If Major = MIS and St_Gender = M and St_HS_Major = Literary and St_Age _group = B then CGPA = A.
5	If Major = Manag then CGPA = A

4.Discussion and findings

The J48 algorithm produces a decision tree with size 25 nodes and number of leaves 16 and the time taken to build model is 0.05 seconds. The analysis of the selected classifier, J48, on the student's dataset shows that the Accounting major is suitable according to the three rules and MIS and the Management major is suitable for any student through one rule as showed in *Table 4*. Finally, we recommend the use of the proposed model in advising students in selecting a suitable major and to identify the right track in higher education which may enhance students' academic performance and decrease dropout.

5.Conclusions and future directions

In this paper, a case study has been presented that shows how advising students in selecting suitable academic major early plays a vital role in any student's life and very important for any higher educational system. Three data mining classification technique were used on real dataset representing students' records in a managerial higher institute in Giza Egypt. We aimed at using the knowledge extracted from the student's data base for giving student advice to select suitable academic majors in the first academic year to enhance decision making for students and educational systems. The implementation of data mining classification techniques was applied by the WEKA data mining tool and the results are reported. In the process of evaluation of the classifier models J48 showed better results with the best five rules that we obtained from 16 rules after eliminate rules that have lower grades. The accuracy and the classification error obtained are 87.64% and 12.36 respectively. So it is selected as the classifier on which we built our proposed student's advising model. For future work, we will generalize the study and add more attributes related to the student's academic qualifications, and apply the model to other universities and institutes in the private education sector. We will extend the experiment using other data mining tools.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Hand DJ. Principles of data mining. Drug Safety. 2007; 30(7):621-2.
[2] Mahboob T, Irfan S, Karamat A. A machine learning approach for student assessment in E-learning using quinlan's C4.5, naive bayes and random forest algorithms. In international conference on multi-topic. 2016 (pp. 1-8). IEEE.
[3] Christian TM, Ayub M. Exploration of classification using NBTree for predicting students' performance. In international conference on data and software engineering 2014 (pp. 1-6). IEEE.
[4] Kumar SV, Padmapriya S. An efficient recommender system for predicting study track to students using data mining techniques. International Journal of Advanced Research in Computer and Communication Engineering. 2014; 3(9):7996-9.
[5] Aulck L, Velagapudi N, Blumenstock J, West J. Predicting student dropout in higher education. ICML workshop on #Data4Good: machine learning in social good applications, New York, USA 2016 (pp. 16-20).
[6] Abu-Oda GS, El-Halees AM. Data mining in higher education: university student dropout case study. International Journal of Data Mining & Knowledge Management Process. 2015; 5(1):15-27.
[7] Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2016.
[8] Lakshmi Devasena C. Comparative analysis of random forest, REP tree and J48 classifiers for credit risk prediction. In international conference on communication, computing and information technology 2014 (pp. 30-6).

Mohamed Hegazy Mohamed is a Teaching Assistant in October 6 University Giza, Egypt. He did his Bachelor of Management in Information System from El Shorouk Academy, Cairo, Egypt. His current research interests are Database, Data Mining, E-commerce, E-business Systems and System Analysis and Design.
Email: mhegazy90@hotmail.com

Hoda Mohamed Waguih is an Associate Professor in the department of Computer and Information Systems Sadat Academy for Management Sciences, Cairo, Egypt. His current research interests are Data Mining, Machine Learning, Predictive Systems and Database.
Email: hoda.waguih@gmail.com