Journal of Computer Science

Original Research Paper

# A Data Mining Model for Students' Choice of College Major Based on Rough Set Theory

**Luai Al-Shalabi**

*Information Technology and Computing, Arab Open University, Al-Ardia 92400, Kuwait*

*Article history:*
Received: 13-06-2019
Revised: 18-07-2019
Accepted: 20-08-2019

Email: lshalabi@aou.edu.kw

**Abstract:** Literature is focusing on identifying factors that influence students' initial choice of major and few have studied students' involvements after registration in a selected major and this study is one of the few. This study aims to determine the important factors that influence high school students' choice of major based on data mining techniques. A questionnaire was designed to collect data from students in different universities in Kuwait and in different faculties such as science, literature, medicine and engineering. Rough set theory for feature selection was used to highlight and explain the significant factors related to students' skills and preferences awareness as well as their experience reflection that are responsible for the development of their satisfaction with the choice of their university majors. The findings of the study revealed that the calculated reducts have a significant influence on the students' choice of the university and collage major. This research contributes to literature by identifying the relationship between the conditional factors of the reduct (also known as the independent variables) and the classification attribute (also known as the dependent variable). The results of the study give valuable information to the high school students so they know the best majors which suite their skills, preference and experiences. This research also help students not to continually change their major because of the wrong choice of major they made which accordingly lead them to dissatisfaction of their major.

**Keywords:** Rough Set, Reduct, University and College Major, Classification

## Introduction

There are thousands of high school students in Kuwait graduate every year. Most of them will enroll in government and private universities in different majors based on their choice as well as their GPA in the high school certificate.

Many students change their major after the first semester or at most after the first year they completed because they could not continue in the current major for different reasons such as they did not like the major, they could not get good GPA and/or they find the current major difficult for them to proceed with. Some of the students may change their major multiple times. Students must be aware of university majors and the suitable ones for them which can satisfy their ambition and ability. Since there is big competition in the market and this leads to big competition between graduate students to get job, students' success in their majors is the first criteria for companies to choose between graduates. The simplest presentation of this success is the student's GPA. In this research, the success of students' choice of major (the suitable major for them) based on their university GPA was studied. After that, the significant factors for those students who have good GPA in their major were also studied. Few researchers have studied students' experiences after enrollment in a selected major (Milsom and Coughlin, 2015), This study built a model of good choice of major from the students' experience after enrollment in their selected major (the main technique) as well as other factors from their pre-university education, ability, preferable, ambition and others. The generated model will be applied to high school students who are going to enter universities in order to advise them and choose a suitable majors for them that guarantee their success and that can put them in high rank between all others graduates who compete them for getting a job.

To better understand university majors that high school students choose, different perspectives have been assumed and consequently different analyzing models have been adopted. Many researchers have developed models based on different factors (input attributes) such as income, parental characteristics, gender, academic ability, personality and influence of significant others and desired outcomes (output attributes or classification attributes), such as enjoying coursework and job satisfaction (Paolillo and Estes, 1982; Cohen and Hanno, 1993; Stinebrickner and Stinebrickner, 2009; Zafar 2011; Beffy *et al*., 2012). Others have used an information processing model which continuously watches the college major choice of students who repeatedly update their major as long as more information is received (Altonji, 1993; Arcidiacono, 2004; Arcidiacono *et al*., 2012).

The objectives of this study are to understand the process that high school students go through in choosing their university majors. This study aims to determine the factors that influence high school students' choice of majors. It also aims to build a model of information which is presented as if-then-else rules that helps high school students to choose the university majors which can satisfy them. The model will give students a list of possible matching majors and then it is their decision to choose the best for them.

The significant of this study is due to different reasons including its uniqueness of contribution to research area since it surveys high school students in Kuwait on their choice of a major based on the training set of university enrolled students (not the high school students) and their GPA plus other pre-school factors and then apply the resulted model to high school students. The study builds an if-then-else model which is easy to understand and easy to implement. The model makes short list of possible majors that satisfy the high school students and give them the ability to choose from the list. Finally, this study will minimize the frequently changing of students' university majors and will lead graduates to high level of success in their working opportunities.

Data mining is the process of discovering or extracting information from stored iceberg of data. Rough set theory is one example of the data mining techniques that are used to discover knowledge. It is a mathematical tool to deal with uncertainty (Pawlak and Skowron, 2007). It can provide a tool for discovering relationships between records and decisions. So, the data set can be reduced to get the minimum representation in terms of decision. Rough set theory will be used in this research in order to identify the most important features that influence high school students to determine their suitable college majors. Data mining was used in medicine (Gagliardi, 2011), Business (Battiti and Passerini, 2010), sensor data mining (Ma *et al*., 2011), learning (Al-Shalabi, 2016), crime (Al-Shalabi, 2017) and so on.

## Literature Review

Choosing the suitable major could depend on several skills and interests. Those who have memorizing skills are welling to choose economy, history and other literary majors whereas those who are interested with numbers and calculations are welling to choose engineering, information technology and science majors such as mathematics, physics and others.

According to prior studies, researchers reported some of the important factors that affect the students' decision of choosing their suitable majors. These factors including but not limited to the followings: gender, family background, personal interests, peer influence, availability of jobs and career opportunities. Nauta (2007) stated that individuals try to choose college majors that are related to their skills and interests. Rajabi (1994; Strasser *et al*., 2002) have studied the students' perceptions towards their major and they reported that the students generally decide on a major based on the job market requirements. Hanson (1994) discussed that family members and friends play an important role in the in the students' choice of their majors. Cohen and Hanno (1993) showed that parents, friends and counselors are not generally affecting the students' choice of majors. Sharifah and Tinggi (2013) showed that parents are not affecting the students' choice of majors. Mazzarol and Soutar (2002) concluded that family members, teachers, seniors, agents and peers may influence the students' choice of majors. Macionis (2000) studied the mass media factor and showed that this factor influence students' choice of major. Dynan and Rouse (1997) showed that media and prior achievement are affecting the students' selection to their majors. Other researchers refused this study and concluded that the media and friends have less influence on the students' choice of majors (Pearson and Dellmann, 1997). Linda (2006) Showed that media influence the students' major selection and discussed that media, television, internet, advertisement and others may affect the behavior of students who follow up these channels in order to collect information about universities, majors and courses prior to their enrollment in the university. Sharifah and Tinggi (2013) reported insignificant differences between the students' major selection and other factors including, personal interest, family members, past achievements, peers and media. Didia and Hasnat (1998; Bauer and Dahlquist, 1999) shown in their research that students' personalities have the first priority in choosing their majors. In the same manner, Worthington and Higgs (2004) showed that personality and personal interests are key roles for students to choose their majors. Walstrom *et al*. (2008) discovered that students usually choose their major based on the jobs and incomes. Leppel *et al*. (2001) concluded that ability, gender, financial stability requirement, race and parental occupation significantly affect the students' choice of their majors. Job opportunities, previous academic experiences, requirement policies to enter into the study of the major, courses' characteristics and college or university reputation are the five factors that were examined by Galotti and Kozberg (1987) in order to study their influence on the students' selection of majors. Kim and Markham (2002) showed that good job, career, abilities, interest of running a business and the good income play an important role in the choice of majors undergraduate students make. Rababah (2016) identified the relationship between the independent variables (reputation of the university, personal interests, job prospect, family members and peers and media) and the dependent variable (student's choice of accounting as a major) and showed that reputation of the university is expected to influence students' choices study.

## Methodology

The main objective of this research is to predict the factors which influence students' choice of the university majors in Kuwait Universities. This research basically study the influence of different factors on the success of student's choice of his/her major including the students' skills, interests, experiences and university achievement represented by the GPA score. Three stages were used sequentially to complete this study: Data collection, data preprocessing engine and model generation which has the important factors.

### *Data Collection*

The questionnaire and many interviews were used in the current study to collect data from students in different universities in order to determine the factors that may influence high school students' choice of majors in Kuwait. Two types of questionnaires were conducted: one for scientific and the other for literary tracks. The total of 806 students from Arab Open University and Kuwait University and from different majors was participating in this study. 447 students were in the literary track and 359 were in the scientific track.

### *Data Preprocessing*

Resolving missing data, data coding and feature extraction are the data preprocessing steps used in this research to make the dataset ready for training. Missing data may produce misleading results. Pyle (1999) demonstrates that the representation and quality of data is first and foremost before running an analysis.

Incomplete data is an example of noise in data. Noise in data may affect the accuracy of the dataset. Removing such noise will improve the accuracy of the dataset. Al-Shalabi *et al*. (2006) highlighted some reasons for missing data including the followings: the value is not relevant to a particular case, not recorded, or ignored because of privacy concerns. One of the solutions for missing data is to delete all records that have missing data (Dempster *et al*., 1977). This solution is conducted in this research because number of records with missing data is low. Exactly 47 and 15 records were rejected from the literary and scientific datasets respectively because they were not completed. Consequently, 400 and 344 completed records from the literary and scientific datasets respectively were accepted and used for further processing.

Students' responds to the questionnaire were choices of texts (nominal). Coding is the process of converting all non-numeric data to numeric data. The analysis of textual responds is slow and of less accuracy. To avoid this, coding all students' responds into numbers is the choice.

Feature selection is one important step to fine the most valuable features that influence the classification attributes. For this study, the choice of suitable university major is represented by the classification attribute (decision) and the process of feature selection is the best way to determine the factors that highly influence this decision. Rough set theory of feature section is used and is explained next

### *Rough Set Theory*

Rough set theory which was introduced by Pawlak 1982 is an important theory for classification problems (Han and Kamber, 2001). The theory is powerful in reducing the dimension of the data set by its data reduction technique. The theory is important in discovering data dependencies and in dealing with missing values. Reduction based on rough set theory will be conducted in this study.

Rough set information system is denoted by $S = (U, A, V, f)$, where $U$ is the universe of discourse which is a non-empty finite set of $N$ objects $\{x1, x2, \cdot\ \cdot\ \cdot, xN\}$. $A$ is a non-empty finite set of attributes such that $a$: $U \rightarrow Va$ for every $a \in A$ ($Va$ is the value set of the attribute $a$).

$$V = \cup a \in A\ Va$$

$f$: $U\times A \rightarrow V$ is the information function such that $f(x, a) \in Va$ for every $a \in A$, $x \in U$. The information system can also be defined as a decision table by $S = (U, C, D, V, f)$. For the decision table, $C$ and $D$ are two subsets of attributes. $A = \{C\cup D\}$, $C \cap D = \emptyset$, where $C$ is the set of condition features and $D$ is the decision attributes.

Let $a \in C\cup D$, $P \subseteq C \cup D$. A binary relation $IND(P)$, called an equivalence (indiscernibility) relation, is defined as follows:

$$IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, f(x, a) = f(y, a)\}$$

The equivalence relation $IND(P)$ partitions the set $U$ into disjoint subsets. Let $U/IND(P)$ denote the family of all equivalence classes of the relation $IND(P)$. For simplicity of notation, *U/P* will be written instead of *U/IND*(*P*). Such a partition of the universe is denoted by *U/P* = {*P*1, *P*2,⋯, *Pi*,⋯}, where, *Pi* is an equivalence class of *P*, which is denoted [*xi*]*P*. Equivalence classes *U/C* and *U/D* will be called condition and decision classes, respectively.

### *Lower Approximation*

Given a decision table *T* = (*U*, *C*, *D*, *V*, *f*). Let *R* ⊆ *C*∪*D*, *X* ⊆ *U* and *U/R* = {*R*1, *R*2, • • •, *Ri*, • • • }. The *R*-lower approximation set of *X* is the set of all elements of *U* which can be with certainty classified as elements of *X*, assuming knowledge *R*. It can be presented formally as:

$$R-\left(X\right)=\cup\left\{Ri \mid Ri\in U/R, Ri\subseteq X\right\}$$

### *Positive Region*

Given a decision table *T* = (*U*, *C*, *D*, *V*, *f*). Let *B* ⊆ *C*, *U/D* = {*D*1, *D*2,⋯, *Di*, ⋯} and *U/B* = {*B*1, *B*2,⋯, *Bi*, ⋯}. The *B*-positive region of *D* is the set of all objects from the universe *U* which can be classified with certainty to classes of *U/D* employing features from *B*, i.e.,:

$$POSB\left(D\right)=\cup Di\in U/DB-\left(Di\right)$$

### *Reduct*

Given a decision table *T* = (*U*, *C*, *D*, *V*, *f*). The attribute *a* ∈ *B* ⊆ *C* is *D−dispensable* in B, if *POSB*(*D*) = *POS*(*B*−{*a*})(*D*); otherwise the attribute *a* is *D−indispensable* in *B*. If all attributes *a* ∊ *B* are *D−indispensable* in *B*, then *B* will be called *D−independent*. A subset of attributes *B*⊆*C* is a *D−reduct* of *C*, iff *POSB*(*D*) = *POSC*(*D*) and *B* is *D−independent*. It means that a reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes. In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe.

### *Reduced Positive Universe and Reduced Positive Region*

Given a decision table *T* = (*U*, *C*, *D*, *V*, *f*). Let *U/C* = {[*u'*1]*C*, [*u'*2]*C*,⋯, [*u'm*]*C*}, Reduced Positive Universe *U'* can be written as:

$$U'=\{u'1,\ u'2,\cdots,\ u'm\}.$$

and:

$$POSC(D)=\left[u'\ i1\ \right]C\cup\left[u'\ i2\ \right]C\cup\cdots\cup\left[u'it\right]C$$

where, ∀*u'is* ∈ *U and* |[*u'is* ]*C/D*| = 1 (*s* = 1, 2,⋯, *t*).

Reduced positive universe can be written as:

$$U'pos=\{u'i1,\ u'i2,\cdots,\ u'it\}.$$

and ∀*B* ⊆ *C*, reduced positive region:

$$POS'B(D)=\cup X\in U/B\wedge X\subseteq U'pos\wedge\left|X/D\right|=1X$$

where, |*X/D*| represents the cardinality of the set *X/D*. ∀*B* ⊆ *C*, *POSB*(*D*) =*POSC*(*D*) if *POS'B* = *U'pos* (Xu *et al*., 2006). It is to be noted that *U'* is the reduced universe, which usually would reduce significantly the scale of datasets. It provides a more efficient method to observe the change of positive region when we search the reducts. No need to calculate *U/C*, *U/D*, *U/B*, *POSC*(*D*), *POSB*(*D*) and then compare *POSB*(*D*) with *POSC*(*D*) to determine whether they are equal to each other or not. We only calculate *U/C*, *U'*, *U'pos*, *POS'B* and then compare *POS'B* with *U'pos*.

### *The Model and the Classifiers*

Two different classifiers were used in order to build a data mining model from the processed dataset. Each of them generates a model with specific accuracy. The model with highest accuracy is recommended. The models are: Decision rules and Decomposition tree and are explained below:

- Decision rules classification (DR) is based on the relationship between conditional values and some prediction where a rule is a context dependent relationship. Rules typically take the form of an (IF-THEN) expression. Decision rules make it possible to classify objects.
- Decomposition Tree (DT) is a supervised learning technique that builds a tree of nodes, leaves and branches. Nodes represent conditions that test the value of the feature. Leaves represent classes of the classification model. Branches represent the chances of features which stop you at the classes. To construct a tree, a top down move is applied until some stopping criterion is met and different methods, such as Gain in entropy, is used for making nodes (Kumar and Chadha, 2011).

## Data Analysis

For data analysis, rough set theory, descriptive statistic and classification accuracy were used. The GPA factor is the key to determine the students who achieved high marks in their major which represents how much correct is their choice to enroll that major. When the accepted records were studied, conclusion was made that there are insignificant results represented by big volume of samples. Those samples are of low and medium GPA which mostly represents the failure of students in choosing their suitable university and college major. In the literary dataset, 17% of the samples have a GPA score less than or equal to 2 points whereas 38.25% of them have GPA score between 2 and 3 points exclusively. In the scientific dataset, 11.63% of the samples have GPA score less than or equal to 2 points whereas 41.57% of them have GPA score between 2 and 3 points exclusively This is a strong indication which shows that there is really big problem in choosing the university and college majors. The hypothesis here is that if the student achieves high marks without difficulty then he/she is somehow satisfied with his/her major, otherwise he/she will change it to another one. Reference to the previous discussion, only rows of high GPA scores (between 3 and 4 inclusively) should be processed further and all other rows will be removed. After that, the GPA factor was removed from the data set because it is no more needed.

Rough set feature selection technique (reduction process) has been applied to the scientific and literary datasets in order to remove the redundant and irrelevant features. The literary dataset consists of 23 factors (questions) and the scientific data set consists of 39 factors. The reduction process minimized the dimension of the literary dataset to 11 factors including the classification factor which is the major whereas it minimized the dimension of scientific dataset to 8 factors including the classification factor which is also the major.

The reduct of the literary dataset was (Q1, Q4, Q7, Q10, Q11, Q12, Q13, Q14, Q20 and Q22) plus the decision or classification factor (Major). The reduct of the scientific dataset was (Q3, Q4, Q12, Q19, Q26, Q27 and Q38) plus the decision or classification factor (Major). Those questions are the most important ones that may affect high school students in the choice of their university and college majors.

Different classifiers were used to test the reliability and consistence of the generated reducts which represent the most important factors that affect the students' choice of majors. The classification model which is represented by if-then rules was generated from the reducts. Results will be discussed in the next section.

## Results

Rough set reduction concepts were used to show the power of the relationship between the classification attribute (major) and the conditional attributes. This process concluded with the required reducts that present the significant factors which are the main keys of choosing the university and college major that suite any student. Classifiers will then be used to check how correct is the generated reducts. This is shown by calculating the accuracy and the coverage of each classifier.

The data mining classification model for the choice of the university and college majors in Kuwait was built from the reduct. It is now able to determine the most match major for high school students which allow them to continue in that major successfully and safely without the need to change this major after some time. Decomposition tree and Decision rules are some of the well-known classification techniques and they were used in this work. The accuracy performance, the coverage of each classifier and the sensitivity metrics were calculated. The accuracy performance is the significant measurement over the coverage. The model of higher accuracy performance would be used to expect the suitable university or college major for any high school student in Kuwait.

The accuracy is defined as the percentage of the instances that are classified correctly by the classifier whereas the coverage is the ratio of classified objects from the class to the number of all objects in the same class. If two classifiers have the same accuracy performance then the model with higher coverage ratio is chosen. The true positive rate which is also called sensitivity, recall, or probability of detection describes the accuracy of the positive cases and gives indication about the power of the classifier.

Table 1 shows the results (accuracy performance and the coverage ratio of each classifier) for the literary data set whereas Table 2 shows the results for the scientific data set. For the literary data set, Decomposition tree and Decision rules are both have accuracy performance (100%) The coverage of the Decision rules is (100%) whereas the coverage of the Decomposition tree is (94.4%). The best choice of classifiers for the literary dataset will be the Decision rules since it has 100% of accuracy performance and coverage ratio. On the other hand and for the scientific dataset, the accuracy performance of the Decision rules is (88.2%) and it is (87.3%) for the Decomposition tree. The coverage of the Decision rules is (100%) whereas the coverage of the Decomposition tree is (88.2%). Results showed that the Decision rules classifier is the best classifier since its accuracy performance is the highest. It also has the coverage ratio (100%) which can classify all the records in the training set. Figure 1 represents the accuracy and the coverage of the literary dataset whereas Fig. 2 represents the accuracy and the coverage for the scientific dataset.

Table 3 is another important table that represents the classification accuracy of each class (major) in the literary dataset given by DR and DT classifiers. Both classifiers are pioneer and are able to classify all tested examples correctly.

Table 4 shows that Decomposition tree classifier is able to test all examples in the dataset whereas decomposition tree is not. Decomposition tree is not able to test 6.7%, 11.8%, 4%, 13.3% and 3% of the

Literature, Business admin, Social science, Law and Islamic studies instances respectively. Combining the accuracy and the coverage metrics, we can say that decision rule classifier has 100% accuracy for the 100% of the, for example, Literature examples (i.e., 100% accuracy for all the Literature examples) while decomposition tree has 100% accuracy for the 93.3% of the Literature examples. As a result of that we can say that decision rule classifier is superior to decomposition tree. Figure 3 represents these results.

Table 5 shows the sensitivity of each major in the literary dataset given by the decision rules and decomposition tree classifiers. Both classifiers are 100% sensitive.

As shown in Table 6, five out of the ten majors that scientific track students enrolled in are from the non-scientific majors denoted by Education, Literature, Business Admin, Social Science and Islamic studies. Both classifiers give closed accuracy with less than 10% difference in suggesting a specific literary major for the high school students. We also notice that both classifiers works well on the other five scientific majors with less than 10% difference in accuracy except for medicine and science majors where decision rule classifier has higher privilege over the decomposition tree classifier Fig. 4 shows the mentioned results.

**Table 1:** The results from the literary dataset

| | Accuracy Performance (%) | Coverage Ratio (%) |
|---|---|---|
| DR | 100 | 100.0 |
| DT | 100 | 94.4 |

**Table 2:** The results from the scientific dataset

| | Accuracy Performance (%) | Coverage Ratio (%) |
|---|---|---|
| DR | 88.2 | 100.0 |
| DT | 87.3 | 88.2 |

**Table 3:** The classification accuracy of each literary dataset major given by DR and DT

| Major | DR (%) | DT (%) |
|---|---|---|
| Literature | 100 | 100 |
| Business Admin | 100 | 100 |
| Social Science | 100 | 100 |
| Law | 100 | 100 |
| Islamic studies | 100 | 100 |

**Table 4:** The coverage ratio of each literary dataset major given by DR and DT

| Major | DR (%) | DT (%) |
|---|---|---|
| Literature | 100 | 93.3 |
| Business Admin | 100 | 88.2 |
| Social Science | 100 | 96.0 |
| Law | 100 | 86.7 |
| Islamic studies | 100 | 97.0 |

**Table 5:** The true positive rate (sensitivity) of each literary dataset major given by DR and DT

| Major | DR (%) | DT (%) |
|---|---|---|
| Literature | 100 | 100 |
| Business Admin | 100 | 100 |
| Social Science | 100 | 100 |
| Law | 100 | 100 |
| Islamic studies | 100 | 100 |

**Table 6:** The classification accuracy of each scientific dataset major given by DR and DT

| Major | DR (%) | DT (%) |
|---|---|---|
| Education | 93.8 | 97.8 |
| IT | 100.0 | 100.0 |
| Literature | 78.6 | 69.2 |
| Business Admin | 80.0 | 75.0 |
| Social Science | 85.7 | 91.7 |
| Islamic studies | 93.3 | 84.6 |
| Engineering | 66.7 | 70.0 |
| Medicine | 100.0 | 75.0 |
| Math | 88.9 | 94.1 |
| Science | 100.0 | 87.5 |

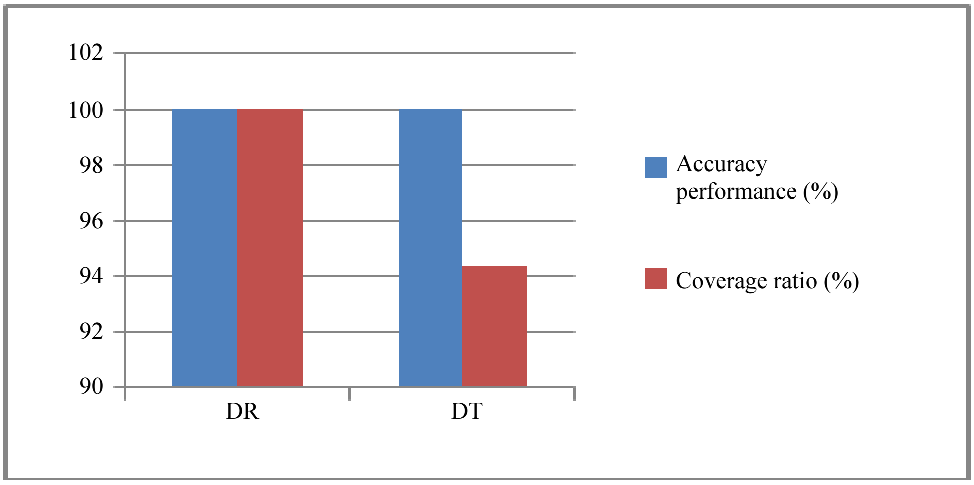

**Fig. 1:** The accuracy and the coverage of the literary dataset

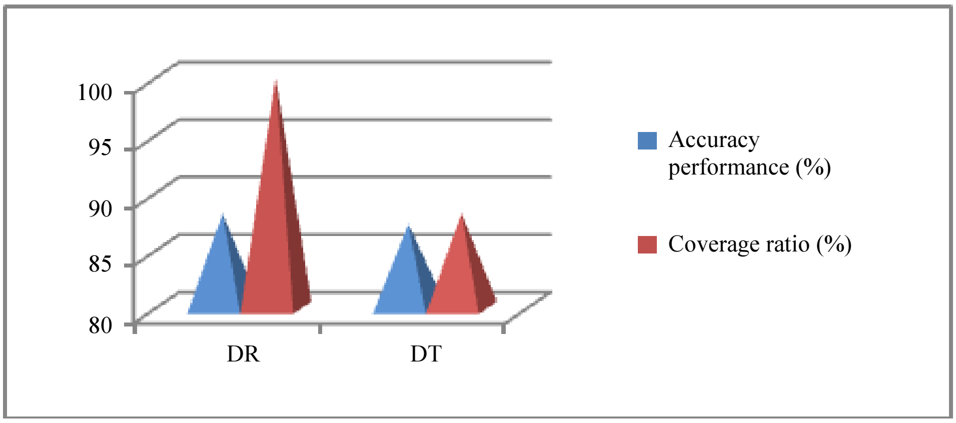

**Fig. 2:** The accuracy and the coverage of the scientific dataset

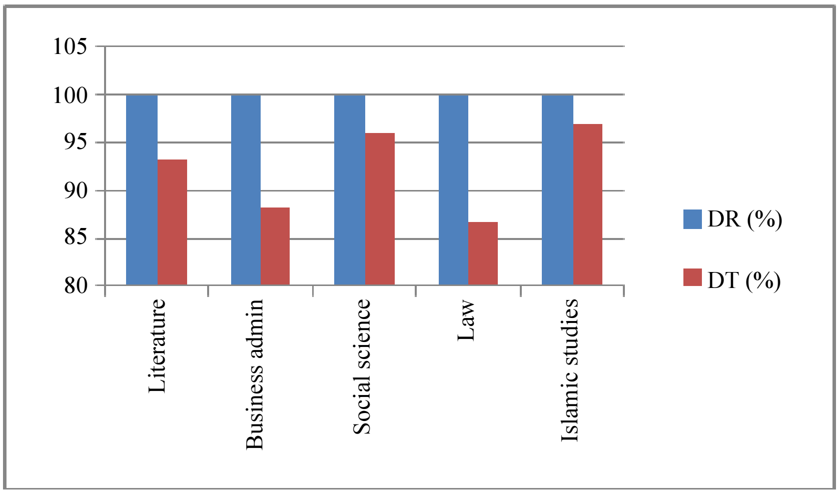

**Fig. 3:** The coverage ratio of each literary dataset major given by DR and DT

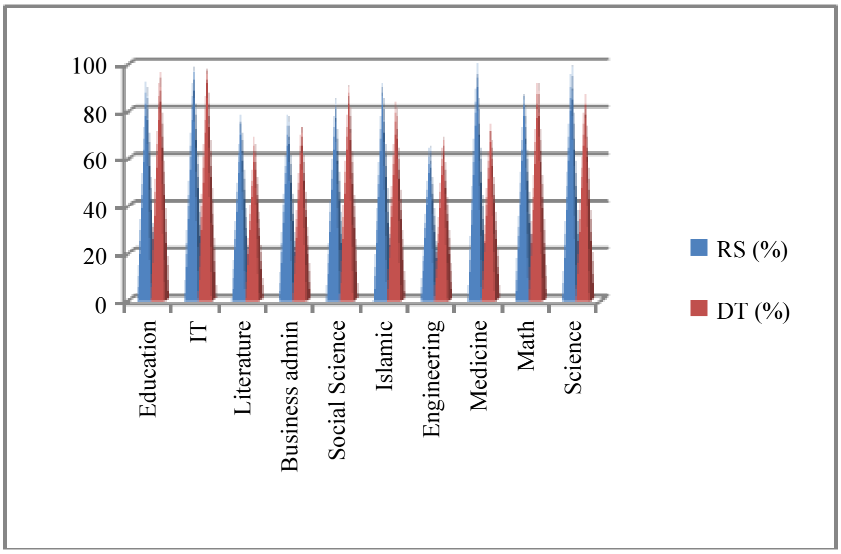

**Fig. 4:** The classification accuracy of each scientific dataset major given by DR and D

Table 7, decision rule classifier is able to classify each example in the scientific dataset to its corresponding major whereas decomposition tree classifier missed that for many examples. For example, decomposition tree classifier is able to classify all the examples of medicine study in the dataset whereas decomposition tree is able to classify 80% of the medicine dataset examples and it could not specify the major for the other 20%. This implies that the decision rule classifier is better than decomposition tree. Figure 5 represents these results.

A perfect predictor would be described as 100% sensitive if, for example, all IT students are correctly identified as IT students which are the correct case for both classifiers as shown in Table 8. Decision rule classifier is 94% sensitive since 94% of Math students are correctly identified by the classifier as Math students whereas decomposition tree classifier is 89% sensitive since 89% of Math students are correctly identified by the classifier as Math students. For the whole system, decision tree sensitivity is 83.8% whereas it is 82.7% for the decomposition tree. Focusing on scientific majors, decision rule classifier is 85.6% sensitive whereas decomposition tree is 82.85 sensitive. For this metric, decision tree is still better than decomposition tree. Figure 6 shows the given results.

Table 9 shows the number of rules given by the decision rule classifier. The classifier generated 172 rules for the literary dataset and 135 rules for the scientific dataset. Figure 7 represents the number of rules generated from the literary and scientific datasets.

Samples of rules generated by Decision rules classifier is given in below:

1. If((attr0=2)&(attr1=3)&(attr2=2)&(attr3=2)&(attr4=2)&(attr5=2)&(attr6=2)) then Class = Education
2. If((attr0=2)&(attr1=3)&(attr2=3)&(attr3=3)&(attr4=3)&(attr5=3)&(attr6=1)) then Class = Education or Class = Islamic Studies
3. If((attr0=1)&(attr1=2)&(attr2=3)&(attr3=2)&(attr4=2)&(attr5=1)&(attr6=1)) then Class = IT
4. If((attr0=1)&(attr1=1)&(attr2=3)&(attr3=3)&(attr4=2)&(attr5=2)&(attr6=1)) then Class = Engineer or Class=Math
5. If((attr0=1)&(attr1=3)&(attr2=2)&(attr3=3)&(attr4=2)&(attr5=3)&(attr6=1)) then Class = Math
6. If((attr0=1)&(attr1=2)&(attr2=2)&(attr3=1)&(attr4=2)&(attr5=2)&(attr6=1)) then Class = Islamic Studies or Class = Engineer

Rules 1, 3 and 5 give one classification value and rules 2, 4 and 6 give two classification values. Rules 2, 4 and 6 give suitable variety of majors that student may choose to study at the university. Suggestions given by the rule are based on the values of the conditional attributes that represents the skills, preference and experience of the student. So the suggestion for the scientific track students could be tertiary or scientific major. A student from scientific track could be distinguished in literary major.

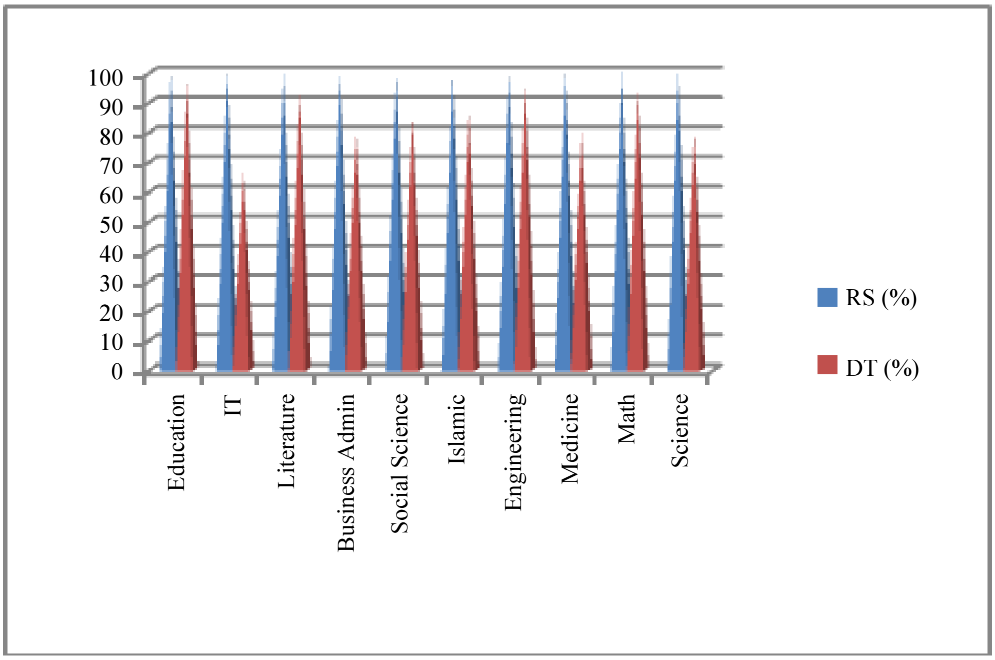

**Fig. 5:** The coverage ratio of each scientific dataset major given by DR and DT

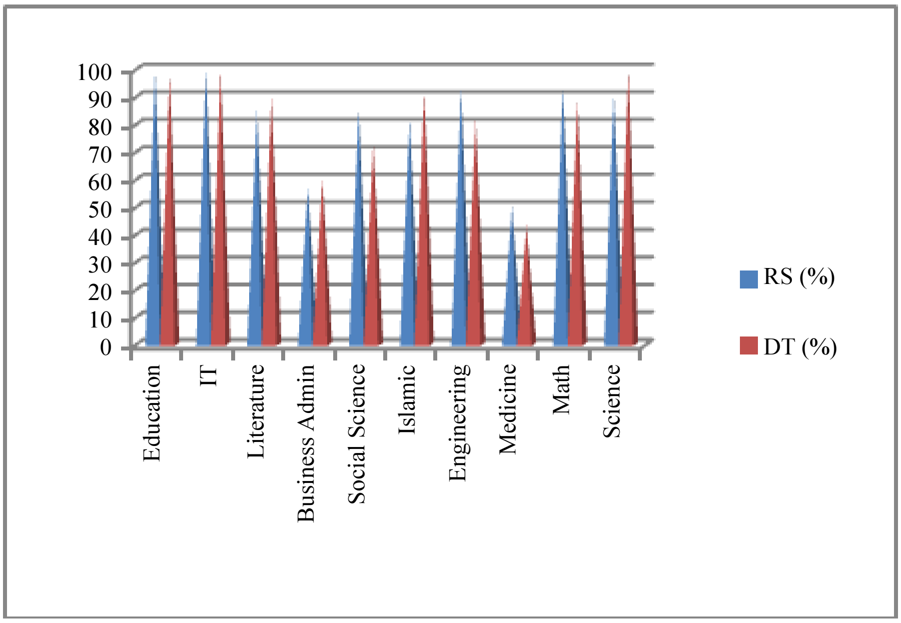

**Fig. 6:** The sensitivity of each scientific dataset major given by DR and DT

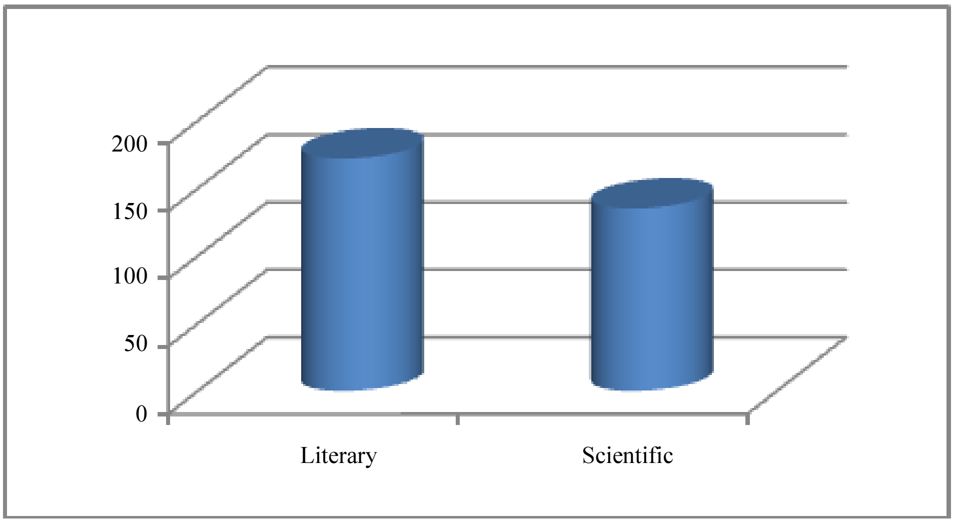

**Fig. 7:** Number of rules given by DR classifier for each dataset

**Table 7:** The coverage ratio of each scientific dataset major given by DR and DT

| Major | DR (%) | DT (%) |
|---|---|---|
| Education | 100 | 95.8 |
| IT | 100 | 66.7 |
| Literature | 100 | 92.9 |
| Business Admin | 100 | 80.0 |
| Social Science | 100 | 85.7 |
| Islamic studies | 100 | 86.7 |
| Engineering | 100 | 95.2 |
| Medicine | 100 | 80.0 |
| Math | 100 | 94.4 |
| Science | 100 | 80.0 |

**Table 8:** The true positive rate (sensitivity) of each scientific dataset major given by DR and DT

| Major | DR (%) | DT (%) |
|---|---|---|
| Education | 100 | 98. |
| IT | 100 | 100 |
| Literature | 85 | 90 |
| Business Admin | 57 | 60 |
| Social Science | 86 | 73 |
| Islamic studies | 82 | 92 |
| Engineering | 93 | 82 |
| Medicine | 50 | 43 |
| Math | 94 | 89 |
| Science | 91 | 100 |

**Table 9:** Number of rules given by DR classifier for each dataset

| Literary | Scientific |
|---|---|
| 172 | 135 |

## Conclusion

This study developed and distributed questionnaires to study the relationship between the conditional attributes denoted by the questions of the questionnaire and the classification attribute denoted by student's choice of university and college majors.

Following conclusion may be drawn from the data mining classification model obtained using rough set approach. From the data mining model and the features reduction analysis results, it is noted that feature selection has an impact on the learning process including the training time, classification time, classification accuracy, coverage ratio and sensitivity as well. Eleven features out of the twenty two predicting features of the literary data set are able to predict new example correctly (100%) by decision rules and decomposition tree classifiers with an advantage for decision rules since it has better coverage percentage value. Also, seven features out of the thirty eight predicting features of the scientific data set are able to predict new example with accuracy of 88.2% and 87.3% by decision rules and decomposition tree classifiers respectively. Results prove that the calculated reducts have the most important predicting features which lead high school students to choose their university and college major successfully.

This article shows that the development of a system which chooses the university and college major for high school students in Kuwait will influence their success during the university study and it will probably minimize the duration of study for those who are expected to change their major if they choose it wrongly from the beginning. This will also improve the universities educational and financial systems since it will minimize the major's transfer and consequently the universities will graduate their students on time (based on the major schedule) which allow them to accept new inputs (students) that will increase the money wise of them.

This research contributed to literature by identifying the relationship between the conditional attributes and the classification attributes (student's choice of university and college major) for both literary and scientific data sets collected from AOU and Kuwait universities in Kuwait. It also shows that only high GPA university students can influence the results given by this research (based on the main technique used).

In near future, other universities will be included in the study as well as other colleges and build comprehensive system for choosing the suitable major for high school students in Kuwait. Other techniques for feature reduction could be used as well as other classifiers. Also, the GPA of the high school certificate will be added to the data sets in order to test its significance on selecting university and college majors for high school students in Kuwait.

## Acknowledgements

The author would like to thank Arab Open University for their support. He also likes to thank many anonymous people for their efforts in improving the readability of this paper including the patience of my wife and kids.

## Funding Information

This research was supported and funded by the research sector, Arab Open University-Kuwait Branch under decision number 19051.

## Ethics

This research article is original and has not been published elsewhere. The corresponding author confirms that there are no ethical issues involved.

## References

Al-Shalabi, L., M. Najjar and A. Al-Kayed, 2006. A framework to deal with missing data in data sets. J. Comput. Sci., 2: 740-745.

Al-Shalabi, L., 2016. Data mining application: Predicting students' performance of ITC program in the Arab Open University in Kuwait-the blended Learning. Int. J. Comput. Sci. Inform. Security, 14: 827-833.

Al-Shalabi, L., 2017. Perceptions of crime behavior and relationships: Rough set based approach, Int. J. Comput. Sci. Inform. Security, 15: 413-420.

Altonji, J.G., 1993. The demand for and return to education when education outcomes are uncertain. J. Labor Economic, 11: 48-83.

Arcidiacono, P., 2004. Ability sorting and the returns to college major. J. Economic, 121: 341-375. DOI: 10.1016/j.jeconom.2003.10.010

Arcidiacono, P., V.J. Holtz and S. Kang, 2012. Modeling college major choice using elicited measures of expectations and counterfactuals. J. Economic., 166: 3-16. DOI: 10.1016/j.jeconom.2011.06.002

Battiti, R. and A. Passerini, 2010. Brain-Computer Evolutionary Multi-Objective Optimization (BC-EMO): a genetic algorithm adapting to the decision maker (PDF). IEEE Transact. Evolut. Comput., 14: 671-687. DOI: 10.1109/TEVC.2010.2058118

Bauer, R. J. and J.R. Dahlquist, 1999. Recognizing and eliminating gender bias in finance education. Financial Practice Edu., 9: 83-91.

Beffy, M., D. Fougere and A. Maurel, 2012. Choosing the field of study in post-secondary education: Do expected earnings matter? Rev. Economics Stat., 94: 334-347. DOI: 10.1162/REST_a_00212

Cohen, J. and D.M. Hanno, 1993. An analysis of the underlying constructs affecting the choice of accounting as a major. Issues Account. Educ., 8: 219-238.

Dempster, A.P., N.M. Larid and D.B. Rubin, 1977. Maximum likelihood from imcomplete data via the Em Algorithm (with discussion). J. Royal Stat. Society, 39: 1-38.

Didia, D. and B. Hasnat, 1998. The determinants of performance in the university introductory finance course. Financial Practice Educ., 1: 102-107.

Dynan, K.E. and C.E. Rouse, 1997. The under representation of women in economics: A study of undergraduate students. J. Economic. Educ., 28: 350-368. DOI: 00220489709597939

Gagliardi, F., 2011. Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction. Artificial Intell. Med., 52: 123-139. DOI: 10.1016/j.artmed.2011.04.002

Galotti, K.M. and S.F. Kozberg, 1987. Older adolescents' thinking about academic/vocational and interpersonal commitments. J. Youth Adolescence, 16: 313-330. DOI: 10.1007/BF02138464

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier, Morgan Kaufman, US, ISBN-10: 0123814804, pp: 744.

Hanson, S.L., 1994. Lost talent: Unrealized educational aspirations and expectations among U.S. Youths. Soc. Educ., 67: 159-183. DOI: 10.2307/2112789

Kim, D. and F.S. Markham, 2002. Why students pursue the business degree: A comparison of business majors across universities. J. Educ. Bus., 78: 28-32. DOI: 08832320209599694

Kumar, V. and A. Chadha, 2011. An empirical study of the applications of data mining techniques in higher education. Int. J. Adv. Comput. Sci. Applic., 2: 80-84. DOI: 10.1.1.631.6325&rep=rep1&type=pdf

Leppel, K., M.L. Williams and C. Waldauer, 2001. The impact of parental occupation and socioeconomic status on choice of college major. J. Family Economic. Issues, 22: 373-394. DOI: 1012716828901

Linda, A., 2006. How to Choose a College Major. 1st Edn., McGraw-Hill.

Ma, Y., Y. Guo, X. Tian and M. Ghanem, 2011. Distributed clustering-based aggregation algorithm for spatial correlated sensor networks. IEEE Sensors J., 11: 641-648. DOI: 10.1109/JSEN.2010.2056916

Macionis, J.J., 2000. Society: The Basics. 13th Edn., Pearson, ISBN-10: 0133752755, pp: 629.

Mazzarol, T. and G. Soutar, 2002. Push-pull factors influencing international students' destination choice. Int. J. Educ. Manage., 16: 82-90. DOI: 09513540210418403

Milsom, A. and J. Coughlin, 2015. Satisfaction with college major: A grounded theory study. NACADA J., 35: 5-14. DOI: 10.12930/NACADA-14-026

Nauta, M., 2007. Assessing college students' satisfaction with their academic. J. Career Assessment, 15: 446-462. DOI: 10.1177/1069072707305762

Paolillo, J.G.P. and R.W. Estes, 1982. An empirical analysis of career choice factors among accountants, attorneys, engineers and physicians. Account. Rev., 57: 785-793.

Pawlak, Z. and A. Skowron, 2007. Rudiments of rough sets. Inform. Sci., 177: 3-27.

Pearson, C. and M. Dellmann, 1997. Parental influence on a student's selection of a college major. College Student J., 31: 301-313.

Pyle, D., 1999. Data Preparation for Data Mining. 1st Edn., Morgan Kaufmann Publishers, Los Altos, California.

Rababah, A., 2016. Factors influencing the students' choice of accounting as a major: The case of X University in United Arab Emirates article. Int. Bus. Res., 9: 25-32. DOI: 10.5539/ibr.v9n10p25

Rajabi, M., 1994. Factors affecting the students accepting in accounting department in the University of Jordan. J., Human. Soc. Sci. Series, 1: 123-158.

Sharifah, S.S. and M. Tinggi, 2013. Factors influencing the students' choice of accounting as a major. IUP J. Account. Res. Audit Pract., 12: 25-42.

Stinebrickner, T.R. and R. Stinebrickner, 2009. Learning about academic ability and the college drop-out decision. Nat. Bureau Economic Res. DOI: 10.3386/w14810

Strasser, S.E., C. Ozgur and D.L. Schroeder, 2002. Selecting a business college major: An analysis of criteria and choice using the analytical hierarchy process. Mid. Am. J. Bus., 17: 47-56. DOI: 19355181200200010

Walstrom, K.A., T.P. Schambach, K.T. Jones and W.J. Crampton, 2008. Why are students not majoring in information systems? J. Inform. Syst. Educ., 19: 43-54.

Worthington, A. and H. Higgs, 2004. Factors explaining the choice of an economics major: The role of student characteristics, personality and perceptions of the profession. Int. J. Soc. Economic., 31: 593-613. DOI: 03068290410529416

Xu, Z., Z. Liu, B. Yang and W. Song, 2006. A quick attibute reduction algorithm with complexity of max(O(|C||U|), O(|C|2|U/C|)). Chinese J. Comput., 29: 391-399.

Zafar, B., 2011. How do college students form expectations. J. Labor Economic. 29: 301-348. DOI: 3a10.1086_2f658091