Intelligent Model for Suitable University Specialization Selection in Palestine

Conference Paper · December 2018

DOI: 10.1109/AICCSA.2018.8612801

3 authors:

Lubaba Tamiza
Ppu
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Ghassan Shahin
Palestine Polytechnic University
7 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Radwan R. Tahboub
Palestine Polytechnic University
35 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Intelligent Model for Suitable University Specialization Selection in Palestine

Lubaba Tamiza
Master of Informatics
Palestine Polytechnic University(PPU)
Hebron, Palestine
Email: lubabatamiza@gmail.com

Ghassan Shahin
IS & MM-CASI
Palestine Polytechnic University (PPU)
Hebron, Palestine Email: gshahin@ppu.edu

Radwan Tahboub, PhD
Dean, College of IT and CE,
Palestine Polytechnic University (PPU)
Hebron, Palestine
radwant@ppu.edu

***Abstract*—Choosing suitable track is a key success in the academic and professional life. Whenever the specialization is appropriate for the student; an increase in students' performance is the natural result. Many studies investigated the influential factors affecting specialization selection by using statistical methods, but none of the researches studied these factors and employed machine learning methods to a develop classification model which can help to choose a suitable specialization.**

In this research, we extracted the local influential factors in our area (Palestine) by using filter approach Correlation-based Feature Selection (CFS) and factor analysis approach Principle Component Analysis (PCA). According to the results, we identified five basic influential factors affecting specialization selection at the universities in Palestine. Then we developed a classification model which might consider the first proposed model studying the influential factors affecting the specialization selection and has the ability to predict the specialization selection for high school students by identifying the suitable specialization based on rules.

A special questionnaire was developed which covers various questions relating the influential factors. Hence, our proposed model depends on extracting the previous knowledge and student experiments. The collected data used as inputs to build our classification model using PART.

According to the results, the accuracy of the proposed model is 77.4% for the training group, and 73.7% for the testing group. The accuracy of the proposed model is 73.7%. The model adopted final 49 rules, which are considered as a map to lead high school students steps toward choosing the suitable specialization.

Keywords: Influential factors, Machine learning, Artificial Intelligent, Principal Component Analysis(PCA), Feature selection.

1. Introduction

Universities specialization selection considered as one of the main decision students can make in their social and career life. The National Career Development Association (NCDA) in the United States pointed out that helping individuals increase self-understanding of their abilities, interests, values, and goals are vital foundations of the career development process [1]. Whenever the specialization is appropriate for the student, performance of the student will be better as a natural result. Choosing a suitable specialization can help students to be creative and more interesting with their study, then the students can complete their high education efficiently. According to [2] noted that unemployment rate related to inefficient university graduates. According to [2], [3], [4], there are many factors affecting university specialization selection. These factors may be related to student characteristics as abilities, interests and other factors.Another study focused on the necessity of compatibility between the career selection and the characteristics as in Hollands theory [5], he supposed that most people are one of six personality types: Realistic, Enterprising, Investigative, Artistic, Social and Conventional. According to a study at Tafila Technical University (TTU), they found that the family and friends, future job factor, career option and occupational prestige considered as important factors in universities specialization selection [3]. As there are influential factors in specialization selection, there are problems which face students in specialization selection.According to [6], found that 25% of students at 12th grade are unable to decide their suitable major career choices. Crites (1969, in Zunker, 1990) [2] suggested that about 30% of the students in high school and college are undecided about a career. This problem continues with the students during the undergraduate period. Moreover, it will affect their performance during their future careers . When students select specialization which does not fit their interests and/or abilities, this may lead to incompetence and lack of qualifications needed by companies and other institutions [7]. This is because the students abilities and orientations are not suited to the course they have taken. Hence, the students career success can be best achieved when the specification is suitable for their personality, ability, intellect and their interests as well choosing a college major [2]. According to [8], many of high school students have failed when they chose the wrong undergraduate program, due to the inappropriate selection decision. In this work, a new proposed model is presented to Palestinian students to select and predict their university specialization better using different influential factors through machine learning algorithm and tools.

2. Literature Review

Many studies contribute to understanding what factors are affecting student specialization selection at the university. Most of the studies examined the reasons which make students select a particular specialization at the university by using statistical methods, but none, if any, used Machine learning ans Statistical methods. A summary of previous work and methods used is presented in table 1.

TABLE 1. LITERATURE REVIEW

Literature Review
Su, Ming-Shang et al. identified four elements, personal factors, career exploration factor, school factor and group factor which wre investigated and analyzed statistically. [4]
Kenneth Eduard et al. determined the factors that affect the high school students in choosing their college courses by using statistical analysis. They concluded that the financial stability of the family is the most factor affected [9].
Pascual determined the influential factors in selecting favorite specialization at the Philippine University. The data were gathered using a standard questionnaire (BOPI). The data were statistically analyzed by SPSS. [2]
Sarwar & Masood determined what the factors affected in choosing a business specialization at the university. According to their statistical study, six factors defined to contribute in specialization selection decision making, which are academic factors, social capital factors, future prospect factors, human capital factors [10].
Al-Rfou, studied the influence of future job factors and personal factors in selecting a business major. It aimed to examine the relationship between students gender, branch in general secondary, general secondary average and the choice of business as a specialization. [3].
Reason, determined the student's demographic variables to predict their retention study as age, gender, ethnicity, color, and high school academic achievement [11].
Kartika, designed self-counseling and career development programs to improve students problem-solving skills [6].
Hollands supposed six model environment that can be interpreted to career orientation - realistic, investigative, artistic, socially enterprising and conventional [5].
Jirapanthong, proposed a classification model for selecting an undergraduate program by using data mining tool WEKA, This model aimed to help undergraduate students in selecting a suitable major [8].
Al-Radaideh et al. used data mining technique to help high school students in choosing their study track at school by analyzing the experience of previous students in the same academic achievement. [12].
In [13], proposed data mining- educational analysis model call (DMEDU). The model was presented as a guideline for the higher educational system to improve decision-making processes as planning, evaluation and consulting [13].
Kovacic examine socio-demographic variables as (gender, education, age, work status, ethnicity, and disability) and course program in what extent that affect on student pass and fail during collage by using data mining methods and statistical analysis [14].
Reddy et al. identified the priorities of vocational education courses as well determined the required courses and less required by using the statistical analysis by studying the relationship between the vocational education and undergraduate students background as student gender, caste, course of study, year of study and the living area [15].
Cortez et al. predicted the secondary students grade of two classes by using the previous students school grades, demographic and social students data. To achieve that they used four data mining techniques which are Decision tree [16].
Sacin, Cesol et al. proposed recommendation system to support students to select the suitable courses and the number of courses. They depended on the real data of students considered as attribute as courses enrolled in, name of course, accumulative GPA, term start, grade. To build this system, C4.5 decision tree was used. [17]

From the previous studies, we concluded that we should study students characteristics, their abilities, interests as well as their living conditions before choosing their specialization.

3. Identifying the Influential Factors in Specialization Selection

Influential is a cognitive factor that tends to have an effect on what you do. Factor: anything that contributes causally to a result [18], we make a survey to find the most significant factors that influence in specialization selection at the university as illustrated in table 9.

4. Methodology

Many of researches covered one or more sides of our study. But there is no research used our methodology to solve the specialization selection problem. The bulk of the studies was exhaustive to find the influential factors affected in specialization selection by using statistical analysis as in [4], [10], [19], [23]. Table 4 illustrates other framework contributed in solving specialization selection. In this research, researcher studied the factors theoretically. Based on that, a special questionnaire was built to investigate these factors in our area, especially in Palestine. Then input data were coded (questionnaire) and analyzed to build the classification model. In contrast, some of the research as in [8], [12] employed machine learning tools and methods to build classification model using decision tree (C4.5 classifier). Both of the previous proposed classification models are good models because they are achieved the desired goals. But in this research may will not achieved our desired goals when using their methodologies because the number of features and the data set are different and the used classification algorithm (C4.5) do not perform our goals which identify the specialization selection for high school students and identify the influential factors in specialization selection. To achieve that we need rule-based classification algorithm as PART algorithm. Table3 illustrates the dataset and the results of the classification models.

Table 4 illustrates the other methodologies to contribute to solving specialization selection problem.

TABLE 2. THE INFLUENTIAL FACTORS

1-Financial factors (Income, Economic status) In some societies considered the financial income as a factor in specialization selection at the university. Each of [2], [10] researches whether if the financial factor has an important impact with students specialization selection. [2], [10]
2- Social factors (Family, peers, parents) [3], [2], [10] [19]
3- Grades and Academic achievement In [8], [12] both of them adopted a Grade Point Average (GPA) as parameters in their classification model for helping high school students in the university specialization selection. [8], [12], [10]
4- Family Background factors (Parents educational degree, Parents occupational classification, Parents socioeconomic status) Parents occupational degree, especially father occupational degree considered as an important factor in specialization selection since father considered as a provider with financial needs at the family [4], [2]
5- Future Prospect Factors future earning, career option, occupational prestige and type of work) [3], [2], [10], [5]
6- Demographic factor (Gender) According to Osakinle and Adegoroye, (Hagedorn, Nora, Pascarella, 1996; Leslie and Oaxaca, 1998; National Research Council, 1991) identified that gender is one of the factors affected in major choice at university. [1], [5]
7- Personal factors (Interests, Talents, Mental abilities, Intellectual skills) [2], [10], [20], [9], [6], [21], [22]
8- University factors (Location, Academic program, Financial aids, Reputation) It is considered as an important factor in our Arab environment. [9], [23]

TABLE 3. THE DATASET AND THE RESULTS FOR OTHER FRAMEWORKS

The classification model	Classification model (Al-Radaideh Q et al.,2011)	Classification model (Jirapanthong. W, 2009)
Number of students	218	4000
Number of features	one feature	seven features
Number of Classes	6 class label	22 class label
Accuracy	87.9%	87.9%

TABLE 4. OTHER FRAMEWORKS TO SOLVE SPECIALIZATION SELECTION PROBLEM

Frame work model	Summary	Method
Classification model (Jirapanthong. W, 2009)	Found classification model for selecting an undergraduate program.	Decision tree- C4.5 algorithm
Classification model (Al-Radaideh Q,2011)	Built classification model for helping high school students choosing the suitable branch at Tawjihi.	Decision tree- C4.5 algorithm
Career planning program (Ming-Shang Su, et al.,2016)	Identified factors affecting the Students career decision-making of Junior High School Students.	Statistical analyzing
Career Planning Program (Pascual. N, 2014)	Identified Factors affecting the Students Career Decision-Making for high school Students.	Statistical analyzing

4.1. Data Collection

A questionnaire was used for data collection, the questionnaire was in two forms hard form and soft form.

4.1.1. Designed questionnaire. Our questionnaire designed based on previous literature which studied the factors and parameters affecting with students when they have chosen their specialization as mentioned in the previous section. Hence, the investigation is compatible with the influential factors that affecting students in their specialization. It was tested and validated by experiments and specialist. To ensure questionnaire distribution among studying sample, an electronic questionnaire was designed using Google forms [24] in addition to the hard copy questionnaire. Each of them contains several parts. Hence, each student answers the questions based on two criteria. First, their studying branch at the high school as (Scientific, Literary, Industrial or Commercial). Secondly; their specialization at the university. The questionnaire basically consisted of three parts, profile part, the second was specialization questions, and the third was general questions. Each student had to go through all parts.

4.1.2. Sampling. In this research, population samples were identified by using the random sample without replacement method [25]. In the data collection process, a random hour was chosen for the questionnaires to be distributed during the lecture for all specialization in Palestine Polytechnic University and Hebron University.

4.2. Data set

The number of students at Hebron University is around 9000 students [26]. Also, the number of students at Palestine Polytechnic university is around 6000 students [27] till the time of data collection. Our questionnaire was distributed to all college students in both universities randomly [25] for students of 2nd, 3rd, 4th and 5th years. Hence, the sample population size around 15000 samples. In our research, we distributed around 700 questionnaires in hard form. In addition to electronic questionnaire form. But the number of valid instances was 700 samples. In this research, we adopted their responds since their Cumulative Grade Point Average (CGPA) over than 65%. Hence, we supposed they are passed in their specialization.

According to data collection by questionnaire, one feature was gotten (attribute) to each question, considered as inputs in the classification model. Hence, the whole number of features is 52 features, and there were 7 classes labels (major), each major contains a sample of students as illustrated in Figure 1.

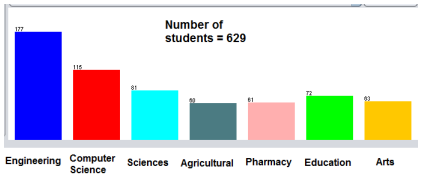

Figure 1. Number of students in each major

4.2.1. Data Cleaning. In our approach, we used the arithmetic mean for filling the missing values to improve the accuracy of our proposed model. The reason to filling the missing values because our questionnaire designed for different specialization and different branches as well each specialization contains different questions than others. Hence, will not answered all the questions and will be as missing values. Assume we have the following values x_1, x_2, $x_3 ... x_n$ the arithmetic mean is calculated as follows.

$$\overline{x} = \sum \frac{x}{n} \tag{1}$$

4.3. Identifying the influential factors in specialization selection

Identifying the influential factors affected in specialization selection based on the previous studies is the basis of our research. The main influential factors were displayed in details in the literature review section and we shall briefly present it as show in 2.

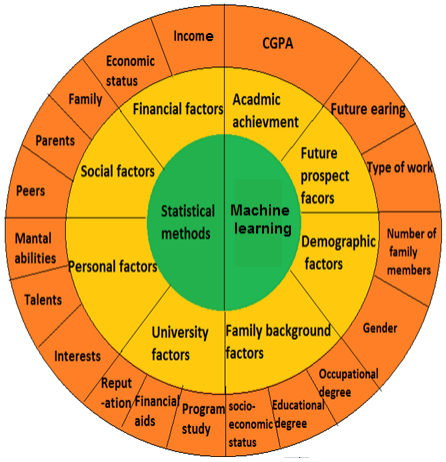

Figure 2. The basic influential factors

4.4. Identifying the Local Influential Factors in Specialization Selection

In order to find the local influential factors, we used reducing dimensionality methods which are feature selection and feature extraction. Reducing the influential factors can help in understanding the most important factors influencing in specialization selection in Palestine. In feature selection, we use filter method which is Correlated based Feature Selection (CFS) [28]. In addition, and to use the extracting feature, Principle Component Analysis (PCA) was used [29]. (PCA) can find a set of correlated factors among, which lead to more general factors (Components) not correlated with each other. From this procedure, we can extract and select the important factors that contribute to specialization selection process in our local area (Palestine). The reducing feature process is illustrated in figure 3

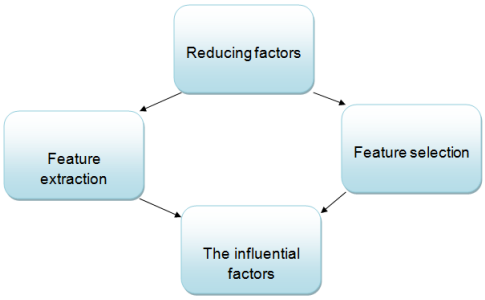

Figure 3. Reducing Features

4.5. Developing our Proposed Classification Model

First, our data sets split into a training set, validation set, and testing set. PART algorithm [30] has been used to build the predictive model, and looking forward to develop a model that has the ability to keep the old one and to contain the old training data by which, the new supplied data for a new student can be used by the old model to choose the suitable specialization. PART algorithm combined between RIPPER and C4.5 classifiers. The importance of using PART classifier is the accuracy and simplicity of produce rules sets by incrementally generated partial decision tree without the need for global optimization. In addition build pruned partial decision tree for a set of instances to produce a single rule; which means this algorithm generates and explores the partial tree to induction one rule. Hence, this decreases the complexity time and improves the performance. PART classifier proposed by Ian Wettin and Eibe Franke. Figure 4 illustrates our methodology for developing the proposed model.

4.5.1. Model Training. This step involves taking the data set which is known by the class label (major) to build and train the classifier and rule extraction. The training set was 629 students of undergraduate students to build the learning classifier.

4.5.2. Model Testing. In our approach we tested the building model in two stages: first, using cross-validation technique to validate and test the building model through the building process. In each time, dividing the training set into k folds. (k-1) folds for training the model and the remaining fold for testing. In this research, we split our data set into 5 folds, 4 folds for learning model represented by gray and 1 for testing model represented by blue, then we measured the performance each time and found the performance average. In the second stage after building the model, we tested it by using a testing set which was not used in the learning model. This set contained 70 students with the unknown class label (major) to predict it. The testing set contains students in the second secondary grade (Tawjihi stage). Figure 5 illustrates the k-fold cross validation, which used in the building model.

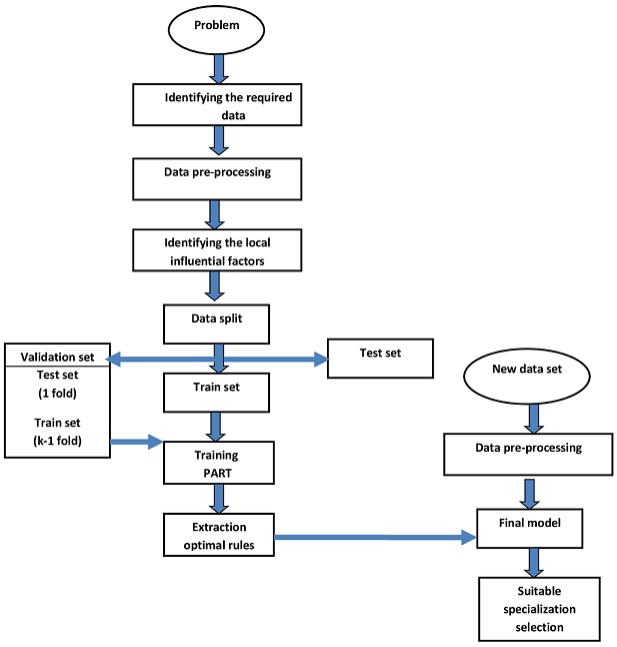

Figure 4. The proposed classification model

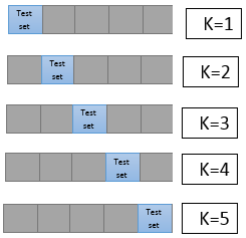

Figure 5. Cross-Validation

5. Results and Findings

5.1. Identifying the Local Influential Factors

Identifying the influential factors in specialization selection in Palestine is one of the main objectives of the research. To achieve this goal we used reducing feature method which contains feature selection and feature extraction [29].

5.2. Identifying the Local Influential Factors Using Feature Selection

Before adopting this method, as a way to find the local influential factors, we tried to use wrapper approach for feature selection, but did not offer the required results because wrapper approach could not handle high numbers of features. But using CFS filter we found it is more suitable and compatible than wrapper as well we get the required results. According to CFS filter and the extraction rules,

we find the main influential factors in our area especially in Palestine as illustrated in the Table 5. In this method the selected features(factors) is arranged according to the merit weight. Thus the feature with the highest merit is considered as the main effective factor in specialization selection. Another approach was used to identify the influential factors which is the statistical analysis by [2] to determine the factors affecting high school students by calculating the mean value and standard deviation of the affecting factors.

TABLE 5. THE LOCAL INFLUENTIAL FACTORS EXTRACTED BY USING (CFS) FILTER

Factor Number	Factor Name
Factor 1	Track
Factor 2	Gender
Factor 3	Skills
Factor 4	Breeding animal at childhood
Factor 5	Tawjihi Average
Factor 6	The reason to choose university
Factor 7	English language average
Factor 8	Problem solving skill
Factor 9	Fixing devices
Factor 10	Driving cars
Factor 11	The preferred subject
Factor 12	Cultivation of plants in childhood
Factor 13	Making scientific experiment at childhood

5.3. Identify the Influential Factors by Using Feature Extraction

We used principle component analysis (PCA) to identify the influential factors. To achieve that, we tested the correlation between the input factors also to test the strength of the relationship between factors thus measure the adequacy of the sample , based on the value of Kaiser Meyer Olkin (KMO). We found the (KMO) value is 0.931, which is an excellent relationship between factors. To find associated factors we depends on the Rotated Component Matrix table after tuning the parameters. We considered the accepted loading values $>= 0.4$. The number of factors were identified according to Eigen -value. Then we used (PCA) because it is considered as one of the most accurate and popular methods of mathematical analysis.

Table 6 illustrated the influential factors affecting specialization selection using (PCA).

TABLE 6. THE EXTRACTED FACTORS BY USING (PCA)

The Factor Number	The Factor Name
Factor 1	Academic achievement
Factor 2	The preferred subjects
Factor 3	The character you cast at childhood
Factor 4	Father career
Factor 5	The reason to choose your university
Factor 6	Reason to choose your major

5.4. Identifying the Final Local Influential Factors by Using (CFS) and (PCA)

The tables 6,5 illustrate the influential factors. The first influential factor is grade and academic achievement, is represented by Tawjih average and the track at the secondary stage this factor was identified by [8], [12] The second influential factor is the demographic factor which represented with the student gender. This influential factor was identified by [1], [5], [10]. Further more our proposed model proved this factor as affecting factors in specialization selection. The third influential factor is the personal factor which is represented by the student interests, skills, talents and mental abilities, it was studied and identified by [2], [6], [10], [21]. The fourth influential factor is family background factors as in [2], [4], affected by parents educational degree, parents occupational classification and parents socioeconomic status. According to our results, family factor represented by father career, which affected specialization selection in Palestine.

The fifth influential factor is university which was identified by [9], [23] which represented with the reasons to choose the university as reputation, place of the university, financial aids, and availability of academic programs. According to our results, we proved that factor an influential factor in specialization selection in Palestine. Figure 6 illustrated the final local influential factors in Palestine.

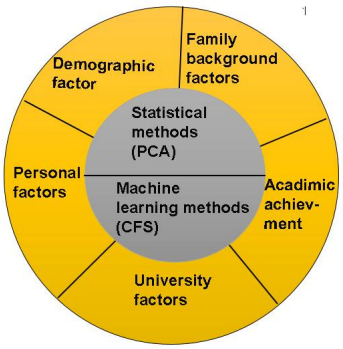

Figure 6. The basic local influential factors in Palestine

6. Developing the Proposed Classification Model

We used PART classifier through WEKA environment and tuning the parameters throw WEKA to obtain optimal classification model. k-fold cross validation was used for evaluation to ensure each training set (k-1) times and one for testing. A cross-validation is applied to evaluate performance predictive model, 4 folds for training the model and 1 fold for testing. By applying (5 folds) cross-validation the accuracy of our proposed model was improved as shown in the Table 7. Our proposed model was built based on students instances at the undergraduate level, who belong to

all university years, from the first year to fifth-year students. This sample of students instances used as a training data to teach and train our classification model. We found 628 valid students instance. And 7 class label (major). The accuracy of the development model was 67.7% when splitting training data into 5 folds using cross-validation. We improved the accuracy to 72.1% when splitting training data into 9 folds. Then we make experiments by reducing the number of classes (majors) and tuning the cross-validation into 5 folds as well we used filters to get rid of the irrelevant features. The accuracy improved to 77.5%. According to the experiments and results, we noted that when reducing the number of the classes, the accuracy will be improved.

TABLE 7. THE ACCURACY OF THE DEVELOPMENT MODEL VS NUMBER OF FOLDS

Number of folds	Accuracy
3 folds	76.9%
4 folds	73.2%
5 folds	77.4%
6 folds	74.8%
7 folds	75%
8 folds	73.9%
9 folds	76.1%
10 folds	73.4%

6.1. Rules Extractions

It is represented as a map to help students choosing their future specialization. The extracted rules represented in an understandable language built using PART classifier. After developing our classification model we extracted these rules. The goals of these rules are to understand the conditions of enrollment in specialization at the university. Each rule is a pattern includes a condition or a set of conditions and factors to meet with each other to represent a rule tables 8 through14 illustrate the extraction rules for each major that help students in choosing their specialization in the future.

TABLE 8. THE RULES OF STUDYING ARTS

1- The student can study this specialization if track = literary AND English Language Avg = v.good AND track = literary AND skills = multi- Languages
2- The student can study Arts if track = literary AND skills = different Cultures
3- The student can study Arts if track = literary AND English-Language Avg = excellent
4-The student can study Arts if track = literary AND Tawjihi Avg = 90-94.9
5- The student can study Arts if track = literary AND Tawjihi Avg = 85-89.9

7. Testing our Prediction Model (Model Verification)

After building our proposed model we tested it by using other students instances with an unknown class, they wanted to know the future specialization selection at the university. The number of students was 70. The accuracy testing of the developing model was 73.7%.

TABLE 9. THE RULES OF STUDYING ENGINEERING

1-The student can study Engineering if Problem-Solving Skill = moderate AND gender = male AND breeding Animal At Child= no AND scientific EXP at Child = chemical Exp AND the preferred Subject = other
2- The student can study Engineering if skills = designing
3- The student can study Engineering if gender = male AND breeding Animal At Child = no AND scientific EXP at Child = chemical Exp AND Problem Solving Skill = moderate AND the preferred Subject=technology AND tawjihi Avg = 80-84.9
4- The student can study Engineering if gender = male AND breeding Animal At Child = no AND scientific EXP at Child = chemical Exp AND Problem Solving Skill = moderate AND the preferred Subject = Scientist Subject
5- The student can study Engineering if the preferred Subject = geography-history AND gender = male
6- The student can study Engineering if track= Scientific AND skills = communication
7- The student can study Engineering if the preferred Subject = English Language AND gender = male
8- The student can study Engineering if the preferred Subject = Arabic Language
9- The student can study Engineering if driving Car = no
10- The student can study Engineering if the preferred Subject = geography-history AND gender = male
11- The student can study Engineering if the preferred Subject = technology
12- The student can study Engineering if fixing Devices = yes

TABLE 10. RULES OF STUDYING AGRICULTURAL

1- The student can study Agricultural if breeding Animal At Child = yes AND scientific EXP at Child = physical Exp
2- The student can study Agricultural if breeding Animal At Child = yes AND scientific EXP at Child = bio Exp AND the preferred Subject = Scientist Subject
3- The student can study Agricultural if scientific EXP at Child = bio Exp AND preferred Major Before = Eng Major
4- The student can study Agricultural if scientific EXP at Child = bio Exp AND the Reason To Choose University = nearness
5- The student can study Agricultural if scientific EXP at Child = bio Exp AND the Reason To Choose University = future Jop Fits
6- The student can study Agricultural if the preferred Subject = health Subj
7- The student can study Agricultural if the preferred Subject = technology AND Tawjihi Avg = 75-79.9

TABLE 11. THE RULES TO STUDYING EDUCATION

1- The student can study Education if track = literary AND skills = analysis-inference
2- The student can study Education if track = literary AND skills = Problem Solving THEN STUDY Education
3- The student can study Education if track = literary AND skills = presentation
4-The student can study Education if track = literary AND Tawjihi Avg = 70-74.9
5- The student can study Education if track = literary

TABLE 12. THE RULES OF STUDYING SCIENCES

1- The student can study Sciences if breeding Animal At Child = yes AND the preferred Subject = Scientist Subject
2- The student can study Sciences if breeding Animal At Child = yes AND the preferred Subject = maths
3- The student can study Sciences if scientific EXP at Child = bio Exp AND preferred Major Before = pharmacy AND breeding Animal At Child = no
4-The student can study Sciences if track = Scientific AND scientific EXP at Child = physical Exp
5- The student can study this specialization if preferred Major Before = IT AND Animal breeding At Child = yes
6-The student can study Sciences if the preferred Subject = maths

TABLE 13. THE RULES OF STUDYING PHARMACY

1-The student can study Pharmacy if the preferred Subject = other AND Tawjihi Avg = 90-94.9
2- The student can study Pharmacy if the preferred Subject = Scientist Subject

TABLE 14. THE RULES OF STUDYING COMPUTER SCIENCES AND INFORMATION TECHNOLOGY

1- The student can study CS &IT if skills = persuasion AND Problem Solving Skill = few
2- The student can study CS &IT if the preferred Subject = management
3- The student can study CS &IT if the preferred Subject = technology AND breeding Animal At Child = no AND skills = persuasion AND gender = male
4- The student can study CS &IT if the preferred Subject = technology AND gender = female AND Tawjihi Avg = 90-94.9
5- The student can study CS &IT if preferred Major Before = sciences
6- The student can study CS &IT if prefer Major Before = arts
7- The student can study CS &IT if prefer Major Before = other AND the Reason To Choose University = good Reputation
8- The student can study CS &IT if prefer Major Before = pharmacy
9- The student can study CS &IT if the preferred Subject = other AND fixing Devices = yes
10- The student can be studying CS &IT if the preferred Subject = technology AND gender = female
11- The student can study CS &IT if the preferred Subject = English Language

8. Performance Measurement

8.1. Confusion Matrix

From confusion matrix results, we noted that in some majors the classification model miss-classified them because some of the majors are similar in the features. Table 15 represents the confusion matrix with n*n size. It is provided with a description of classifying the examples into class labels based on a predictive model.

TABLE 15. THE RESULTS OF CONFUSION MATRIX FOR PREDICTION MODEL

A	B	C	D	E	F	G	Class label
137	24	5	1	10	0	0	A= Engineering
16	85	9	0	5	0	0	B=IT and CE
5	14	32	15	15	0	0	C= Sciences
2	7	23	20	8	0	0	D= Agricultural
5	10	9	7	30	0	0	E= Pharmacy
0	1	0	0	0	62	9	F= Education
1	0	0	0	0	5	57	G= Arts

8.2. Area Under Roc Curve

The area under roc curve makes a description of the examples that classified correctly in the proposed prediction model. Whenever the ROC curve for the classifier is nearest to 1 value, then the average of the AUC is well. Table16 show the results of the AUC for each class (major). The highest value of AUC for Arts major (AUC, 0.986), and the lowest AUC was 0.818 for Sciences major. So this value nearest to 1 value and we can depend on it to predict the suitable major in the future dataset.

Table 18 illustrated the comparison of proposed model before reduced number of majors and after that.

From the results of the Table 18, we found that the accuracy of our model with 10 classes is (ACC,43.2%), (AUC average, 0.788). But with reduced number of classes (majors) to 7 classes; the accuracy of AUC was improved to (ACC, 72.1%), (AUC average, 0.908). While the majors were reduced to 6, the model building became more accurate than before with (ACC, 77.4%) and the performance of AUC was (0.926). Based on our experiments and results, and by reducing the number of classes from 10 classes then to 7 classes then to 6, the ACC and AUC were improved.

TABLE 16. AREA UNDER ROC CURVE FOR THE PREDICATION MODEL USING PART

Class	TP Rate	FP Rate	Precision	ROC Area
Engineering	0.774	0.064	0.825	0.926
IT and CS	0.739	0.109	0.603	0.897
Sciences	0.395	0.084	0.410	0.818
Agricultural	0.333	0.040	0.465	0.829
Pharmacy	0.492	0.067	0.441	0.824
Education	0.861	0.009	0.925	.0974
Arts	0.905	0.016	0.864	0.986
ROC Average				0.899

TABLE 17. CONFUSION MATRIX FOR 6 CLASS LABEL

A	B	C	D	E	F	Class label
151	15	4	7	0	0	A = Engineering
16	81	10	8	0	0	B = IT-CE
4	20	108	9	0	0	C = Sciences
4	10	20	27	0	0	D = Pharmacy
0	0	0	0	62	10	E = Education
0	1	0	0	4	58	F = Arts

9. Comparing Proposed Model with Other Approaches

In this experiment, we are comparing our proposed model with another classifier to validate performance. We used the RIPPER classifier for comparison [31], we found that the extracted rules of RIPPER method were less detailed and contained few features and fewer rules than PART method. Based on the rules, each rule had one repeated idea at each time, which cannot be relied on to be a guideline to help the high school for future specialization selection.

According to the results shown in Tables 19, the classifier PART with (ACC, 77.4%) is better than the classifier RIPPER [31] with (ACC, 71%) for 6 classes (majors). And the ACC of PART for 7 classes is better than RIPPER classifier.

TABLE 18. COMPARING PART RESULTS WITH CHANGING NUMBER OF CLASSES

Number of rules	Number of classes	Accuracy	ROC Curve average
49	6	77.4%	0.926
52	7	72.1%	0.908
76	10	43.7%	0.788

TABLE 19. THE RESULT OF COMPARISON BETWEEN PART AND RIPPER

Classifier	**Number of classes**	**Number of rules**	**Accuracy**
PART	7	52	67.7%
RIPPER	7	27	64.7%
PART	6	49	77.4%
RIPPER	6	20	71%

10. Conclusions

One of the objectives of our research is to identify the influential factors affected the specialization selection based on literature review. Hence, we studied the influential factors and summarized them into the following factors; social factors, future Prospect factors, personal factors, grades and academic achievement, family background factors, financial factors, university factors and demographic factors. Each factor contains sub factors which we studied it in details.

The second objective of our thesis is to identify the local influential factor in Palestine. To achieve this objective, we used reduce dimensionality technique which contains two methods, feature selection, and feature extraction. In the feature selection method, we used the filter Correlated-based Feature Selection (CFS) to identify the important factors which have a significant impact and lower error rates. In feature extraction method, we used principle component analysis (PCA) to find the factors with the significant correlation between other and extract components include the correlated factors. The results of the research argued that there are five influential factors in Palestine which are family background factor, demographic factor, personal factor, university factor and academic achievement factor.

The third objective of our research is developing an intelligent model, which can help high school students to predict what specialization suitable for them in future. The accuracy of the model was 77.4%. Then the model was tested with a sample of students who were not involved in the training set. The accuracy of the model for the testing set was 73.2%, so the developing model was considered to be a good model.

paper.tex

References

[1] K. U. Lazarus and C. Ihuoma, "The role of guidance counsellors in the career development of adolescents and young adults with special needs," *British Journal of Arts and Social Sciences*, vol. 2, no. 1, pp. 51–62, 2011.

[2] N. T. Pascual, "Factors affecting high school students career preference: A basis for career planning program," *International Journal of Science: Basics and Applied Research*, vol. 16, no. 1, pp. 1–14, 2014.

[3] A. N. Al-Rfou, "Factors that influence the choice of business major evidence from jordan," *Journal of Business and Management*, vol. 8, no. 2, pp. 104–8, 2013.

[4] M.-S. Su, T.-C. Chang, C.-C. Wu, and C.-W. Liao, "Factors affecting the student career decision-making of junior high school students in central taiwan area," *International Journal of Information and Education Technology*, vol. 6, no. 11, p. 843, 2016.

[5] S. R. Porter and P. D. Umbach, "College major choice: An analysis of person–environment fit," *Research in Higher Education*, vol. 47, no. 4, pp. 429–449, 2006.

[6] A. Kartika, "The effect of counseling in self and career development to enhance students awareness of their personal and study problems," 2008.

[7] N. Lawrence K. Jones, Ph.D., *Choosing a College Major Based on Your Personality*, 2012 (accessed February 3, 2016), http:https://www.montclair.edu/median.pdf.

[8] W. Jirapanthong, "Classification model for selecting undergraduate programs," in *Natural Language Processing, 2009. SNLP'09. Eighth International Symposium on*. IEEE, 2009, pp. 89–95.

[9] T. B. E. D. M. K. D. E. M. F. G. G. M. S. S. Z. M. K. R. Alba, Kenneth Eduard Castillo Bertol. (2016, Sep.) https://www.coursehero.com/file/12317190/the-factors-that-affect-students-decision-in-choosing-their-college-courses/.

[10] A. Sarwar and R. Masood, "Factors affecting selection of specialization by business graduates," *Science International*, vol. 27, no. 1, 2015.

[11] R. D. Reason, "Student variables that predict retention: Recent research and new developments," *Naspa Journal*, vol. 40, no. 4, pp. 172–191, 2003.

[12] Q. Al-Radaideh, A. Al Ananbeh, and E. Al-Shawakfa, "A classification model for predicting the suitable study track for school students," *Int. J. Res. Rev. Appl. Sci*, vol. 8, no. 2, pp. 247–252, 2011.

[13] M. Beikzadeh and N. Delavari, "A new analysis model for data mining processes in higher educational systems," *On the proceedings of the 6th Information Technology Based Higher Education and Training*, pp. 7–9, 2005.

[14] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data." 2010.

[15] P. A. Reddy, D. U. Devi, and E. M. Reddy, "A study of the vocational education preferences and interests of the indian undergraduate students," *Bulgarian Journal of Science and Education Policy (BJSEP)*, vol. 5, no. 1, pp. 94–114, 2011.

[16] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[17] C. V. Sacin, J. B. Agapito, L. Shafti, and A. Ortigosa, "Recommendation in higher education using data mining techniques," in *Educational Data Mining 2009*, 2009.

[18] http:// www.vocabulary.com. Retrieved at May 12,2016.

[19] N. X. Thao, "Examining family and community influences on the attitudes to education and career aspirations of hmong/mong high school students," Ph.D. dissertation, University of Minnesota, 2009.

[20] http:// http://www.dictionary.com. Retrieved at May 20,2016.

[21] D. Kim, F. S. Markham, and J. D. Cangelosi, "Why students pursue the business degree: A comparison of business majors across universities," *Journal of Education for Business*, vol. 78, no. 1, pp. 28–32, 2002.

[22] D. Q. Nguyen, "The essential skills and attributes of an engineer: A comparative study of academics, industry personnel and engineering students," *Global J. of Engng. Educ*, vol. 2, no. 1, pp. 65–75, 1998.

[23] A. Nora, "The role of habitus and cultural capital in choosing a college, transitioning from high school to higher education, and persisting in college among minority and nonminority students," *Journal of Hispanic Higher Education*, vol. 3, no. 2, pp. 180–208, 2004.

[24] https://docs.google.com.

[25] P. L. Barreiro and J. P. Albandoz, "Population and sample. sampling techniques," *Management Mathematics for European Schools MaMaEusch (994342-CP-1-2001-1-DECOMENIUS-C21*, 2001.

[26] (2016, Sep.) http://www.hebron.edu.

[27] (2017, Feb.) https://www.ppu.edu/p/sites/default/files/ppu%20guide%202016-2017.pdf.

[28] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[29] E. Alpaydin. (2010) Introduction to machine learning. 2-nd edition.

[30] I. H. Witten, "Data mining with weka," *Class Lesson, Department of Computer Science University of Waikato, New Zealand.*, 2013.

[31] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 115–123.