Intelligent Model for Suitable University Specialization Selection in Palestine

Conference Paper - December 2018
DOI: 10.1109/AICCSA.2018.8612801

CITATIONS
0

READS
126

3 authors:

Lubaba Tamiza
Ppu
1 PUBLICATION 0 CITATIONS
SEE PROFILE

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

Ghassan Shahin
Palestine Polytechnic University
7 PUBLICATIONS 4 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Secure Data Sharing Polices and Architecture View project

Adaptive virtual classroom based on students modeling View project

All content following this page was uploaded by Radwan R. Tahboub on 29 December 2018.

The user has requested enhancement of the downloaded file.
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
Hebron, Palestine
Email: gshahin@ppu.edu

Radwan Tahboub, PhD
Dean, College of IT and CE,
Palestine Polytechnic University (PPU)
Hebron, Palestine
Email: 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 develop a 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 (NCDAA) 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 as Statistical methods. A summary of previous work and methods used is presented in table 1.

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 achieved the desired goals. But in this research may not achieve 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. Table 3 illustrates the dataset and the results of the classification models.

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

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.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(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.

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.
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.

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.

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 Kaisser Meyer Olikin (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).

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.

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. 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.

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.

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 14. THE RULES OF STUDYING COMPUTER SCIENCES AND INFORMATION TECHNOLOGY

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>Class label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>137</td>
<td>24</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>A = Engineering</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>85</td>
<td>9</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>B = IT and CE</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>14</td>
<td>32</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>C = Pharmacy</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>7</td>
<td>23</td>
<td>20</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>D = Agricultural</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>E = Computer Science</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>9</td>
<td>F = Education</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>57</td>
<td>G = Arts</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>Class label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>137</td>
<td>24</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>A = Engineering</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>85</td>
<td>9</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>B = IT and CE</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>14</td>
<td>32</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>C = Pharmacy</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>7</td>
<td>23</td>
<td>20</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>D = Agricultural</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>E = Computer Science</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>9</td>
<td>F = Education</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>57</td>
<td>G = Arts</td>
</tr>
</tbody>
</table>

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. Table 16 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.

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 16. AREA UNDER ROC CURVE FOR THE PREDICTION MODEL USING PART

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>0.774</td>
<td>0.064</td>
<td>0.825</td>
<td>0.926</td>
</tr>
<tr>
<td>IT and CS</td>
<td>0.739</td>
<td>0.109</td>
<td>0.603</td>
<td>0.897</td>
</tr>
<tr>
<td>Sciences</td>
<td>0.395</td>
<td>0.084</td>
<td>0.410</td>
<td>0.818</td>
</tr>
<tr>
<td>Agricultural</td>
<td>0.333</td>
<td>0.040</td>
<td>0.465</td>
<td>0.829</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>0.492</td>
<td>0.097</td>
<td>0.441</td>
<td>0.824</td>
</tr>
<tr>
<td>Education</td>
<td>0.861</td>
<td>0.009</td>
<td>0.925</td>
<td>0.974</td>
</tr>
<tr>
<td>Arts</td>
<td>0.905</td>
<td>0.016</td>
<td>0.864</td>
<td>0.986</td>
</tr>
<tr>
<td>ROC Average</td>
<td></td>
<td></td>
<td></td>
<td>0.899</td>
</tr>
</tbody>
</table>

TABLE 17. CONFUSION MATRIX FOR 6 CLASS LABEL

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>Class label</th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
<td>16</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>A = Engineering</td>
</tr>
<tr>
<td>16</td>
<td>81</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>B = IT and CE</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>108</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>C = Sciences</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>20</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>D = Pharmacy</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>E = Education</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>58</td>
<td>0</td>
<td>F = Arts</td>
</tr>
</tbody>
</table>

TABLE 18. COMPARING PART RESULTS WITH CHANGING NUMBER OF CLASSES

<table>
<thead>
<tr>
<th>Number of rules</th>
<th>Number of classes</th>
<th>Accuracy</th>
<th>ROC Curve average</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>6</td>
<td>77.4%</td>
<td>0.926</td>
</tr>
<tr>
<td>52</td>
<td>7</td>
<td>72.1%</td>
<td>0.908</td>
</tr>
<tr>
<td>76</td>
<td>10</td>
<td>43.7%</td>
<td>0.788</td>
</tr>
</tbody>
</table>

TABLE 19. THE RESULT OF COMPARISON BETWEEN PART AND RIPPER

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Number of classes</th>
<th>Number of rules</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PART</td>
<td>7</td>
<td>52</td>
<td>67.7%</td>
</tr>
<tr>
<td>RIPPER</td>
<td>7</td>
<td>27</td>
<td>84.7%</td>
</tr>
<tr>
<td>PART</td>
<td>6</td>
<td>49</td>
<td>77.4%</td>
</tr>
<tr>
<td>RIPPER</td>
<td>6</td>
<td>20</td>
<td>71%</td>
</tr>
</tbody>
</table>
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.

References


