Classification Model Development for Predicting Study Track for the Students of School

MAI Navid[1] ,NH Niloy[2]

Ruhea College, Bangladesh
**Corresponding author's E-mail: niloynh1997@gmail.com*

ABSTRACT:

The article utilized data mining techniques to provide a classification approach to support basic school students in selecting the suitable track. There are a set classification rules that were extracted from the decision tree to predict and classify the class label for each student. The main problem in the selection of an academic track in basic Jordanian schools is the lack of useful knowledge for students to support their planning. One of the most important issues to succeed in the academic life is to assign students to the right track when they arrive to the end of the basic education stage. A confusion matrix is built to evaluate the model where the 10-fold Cross Validation method was used for accuracy estimation of the model. The overall accuracy of the model was 87.9% where 218 students were correctly classified out of the 248 students. For this purpose, a decision tree classification model was developed to determine which track is suitable for each student. Several classification rules were extracted from the decision tree. The overall accuracy of the model was 87.9% which is rated to be very good. The set of the extracted classification rules can be used to build a recommendation system to help the management of schools to best determine the suitable track for the new students who finished the basic education tier.

KEYWORDS: Data Mining, Classification, Decision Tree, Basic Education.

1. INTRODUCTION:

School education in Jordan is a two-tier system; the first tier has ten years of study covering basic education followed by the second tier which has two years of secondary education. The basic education acts as a bridge to the secondary education. The secondary education tier is important because it is a deciding factor for opting desired subjects of study in the higher education level where it acts as a bridge between school education and the higher learning specializations that are offered by colleges and universities. The basic education system is graded from 1 to 10 where after finishing the 10th grade; students are distributed into different academic tracks such as Scientific, Literary, Information Management and others. One of the most important issues to succeed in the academic life is to assign a student to a right track, when they arrive to the end of the basic education stage. The distribution of students depends on several criteria that include the accumulated average of each student, especially in scientific courses, the average marks of the 8th, 9th, and 10th grades, as well as the ratio of the credit average that is accepted to in a specific track. The main problem in the selection of an academic track in basic Jordanian schools is that students are not supported with the required information and analytical information to support their planning. Many students still fail in selecting the right track. This is one of the reasons that lead to a low education quality. To improve the quality of education in Jordan, data mining techniques can be utilized to improve the traditional process that is used to distribute students to right tracks according to their capabilities. Data miming consists of a set of techniques that can be used to extract relevant and interesting knowledge from huge amount of data. It is a technique that can be used to analyze dataset. Data mining technique fall into three methods; which are association rule mining, classification and prediction, and clustering (Han *et al*., 2011). Association rule mining (ARM) is a method for discovering interesting relations between variables in a transaction database. ARM is a famous technique in basket analysis. Classification techniques are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends. There are several algorithms for data classification; one of them is the decision tree classification technique. Generally, this paper is a preliminary attempt to use data mining concepts; particularly classification, to help in supporting the quality of the educational system by evaluating student data to study the main attributes that may affect the student classification in basic school. The paper applies the data mining concepts to develop a model for supporting the selection of academic education track. This work proposes a model based on a data mining technique to help students to choose a suitable track by analyzing the experience of previous student with similar academic achievement. For this purpose we have used the decision tree classification technique to build a model for predicting the suitable track for student when they finish the basic education stage. A set of classification rules was extracted from the generated model.

2. RELATED WORK:

There are several researchers who have focused on educational data mining. One of them is Warapon in (Waraporn, 2009) who presented the use of data mining techniques, particularly classification, to supports high school students in selecting undergraduate programs. Warapon proposed a classification model to give guidelines to students, especially, for the undergraduate programs for making possible better academic plans. The decision tree technique was applied to determine which major is best suitable for students. Al-Radaideh *et al*. (2006) proposed to use data mining classification techniques to enhance the quality of the higher educational system by evaluating students' data that may affect the students' performance in courses. They used the CRISP framework for data mining to mine students' related academic data. A classification model was built using the decision tree method. They used three different classification methods ID3, C4.5 and the NaïveBayes. The results indicated that the decision tree model had better prediction accuracy than the other models. As a result, a system was built to facilitate the usage of the generated rules that students need to predict the final grade in the C++ undergraduate course.

Cesar et al. (2009) proposed the use of a recommendation system based on data mining techniques to help students to make decisions related to their academic track. The system provided support for students to better choose how many and which courses to enroll on. As a result, the authors developed a system that is capable to predict the failure or success of a student in any course using a classifier obtained from the analysis of a set of historical data related to the academic field of other students who took the same course in the past. Naeimeh et al. (2004) have presented and justified the capability of data mining technologies in the context of higher educational system by proposing a new model; called (DM_EDU) that was used as a roadmap for the application of data mining in higher educational system, for improving the efficiency and effectiveness of the higher educational process. This model was used for analyzing the current work of data mining in education and identifying the existing gaps. It also provided an opportunity for researchers to be familiar with the existing area of study for data mining in education. Higher educational institutes can use this model to identify which part of their processes can be improved by data mining technology and how they can achieve this goal. They have used the model for using data mining technology in multimedia university of Malaysia (MMU) educational process and by developing an appropriate data mining system. Pathom *et al*. (2008) proposed a classifier algorithm for building Course Registration Palning Model (CRPM) from historical dataset. The algorithm is selected by comparing the performance of four classifiers include Bayesian Network, C4.5, Decision Forest, and NBTree. The dataset were obtained from student enrollments including grade point average (GPA) and grades of undergraduate students. As a result, the NBTree was the best of the four classifiers. NBTree was used to generate the CRPM, which can be used to predict student class of GPA and consider student course sequences for registration planning. Tissera *et al*. (2006) presented a real-world experiment conducted in an ICT educational institute in Sri Lanka, by analyzing students' performance. They applied a series of data mining task to find relationships between subjects in the undergraduate syllabi. They used association rules to identify possible related two subjects' combination in the syllabi, and apply correlation coefficient to determine the strength of the relationships of subject combinations identified by association rules. As a result, the knowledge discovered can be used for improving the quality of the educational programs. Nguyen *et al*. (2007) compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes. These predictions are most useful for identifying and assisting failing students, and better determine scholarships. As a result, the decision tree classifier provided better accuracy in comparison with the Bayesian network classifier. Muslihan *et al*. (2009) have compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. Students' data were collected from the data of the National Defence University of Malaysia (NDUM). As a result, the technique that gives accurate prediction and classification was chosen as the best model. Using the proposed model, the pattern that influence the student's academic performance was identified. Naeimeh *et al*. (2005) have presented and justified the capability of data mining in the higher education system by offering an enhanced version of (DM_EDU) analysis model. One of the most important parts of the model is "student assessment". To prove the model correctness, authors have implemented one of the sections of the DM_EDU in MMU. As a result, they claimed that the model has improved the quality of the management system. The same authors of Naeimeh et al. (Naeimeh D., Somnuk P., and Mohammad B. 2008) have discussed how the various data mining techniques can be applied to the set of educational data and what new explicit knowledge or models can be discovered. The models discussed are classified based on the type of techniques used, including predictive and descriptive. The obtained rules from each model are translatedto plain English to be used as a factor to be considered by the managerial system to either support their current decision makings or help them to set new strategies and plan to improve their decision making procedures. The main idea of this analysis is organized into the DM-HEDU guideline proposed by the authors, which targets the superior advantages of data mining in higher learning institution. The authors have presented several projects of using data mining in higher education. Ramaswami and Bhaskaran (2010) have constructed a predictive model called CHAID with 7- class response variable by using highly influencing predictive variables obtained through feature selection so as to evaluate the academic achievement of students at higher secondary schools in India. Data were collected from different schools of Tamilnada, 772 students' records were used for CHAID prediction model construction. As a result, set of rules were extracted from the CHAID prediction model and the efficiency was found. The accuracy of the present model was compared with other models and it has been found to be satisfactory. Yiming *et al*. (2000) have presented a real-life application for the Gifted Education Programme (GEP) of the ministry of education (MOE) in Singapore. They have focused only on selecting weak school students for remedial classes based on association rules. Traditionally, a cut-off mark was used to select the weak students who must take further courses. This traditional method requires too many students to take part in the remedial classes. Authors presented new scoring technique; called Scoring Based on Associations (SBA). Three scoring measures namely; Scoring Based on Associations (SBA-score), C4.5-score and NB-score for evaluating the prediction in connection with the selection of the students for remedial classes were used with other input variables like sex, region and school performance over the past few years. It was found out that the predictive accuracy of SBA-score methodology was 20% higher than that of C4.5 score, NB-score methods, as well as traditional scoring methods.

3. BUILDING THE CLASSIFICATION MODEL:

In general, data classification is a two-step process. In the first step, which is called the learning step, a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database instances. Each instance is assumed to belong to a predefined class. In the second step, the model is tested using a different data set that is used to estimate the classification accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data instances for which the class label is not known. At the end, the model acts as a classifier in the decision making process. There are several techniques that can be used for classification such as decision tree, Bayesian methods, rule based algorithms, and Neural Networks.Fadzilah and Abdullah (Fadzilah and Abdoulha, 2009) have presented the results of applying data mining techniques to enrollment data of Sebha University in Libya. Two main approaches were used; descriptive and predictive approaches. Cluster analysis was performed to group the data into clusters based on their similarities. For predictive analysis, three techniques have been used Neural Network, Logistic regression, and the Decision Tree. After evaluating these techniques, Neural Networks classifier was found to give the highest results in term of classification accuracy. Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. Decision tree can produce a model with rules that are human-readable and interpretable. Decision Tree has the advantages of easy interpretation and understanding for decision makers to compare with their domain knowledge for validation and justify their decision. Some of decision tree classifiers are C4.5/C5.0/J4.8, NBTree, and others. The C4.5 technique is one of the decision tree families that can produce both decision tree and rule-sets; and construct a tree for the purpose of improving prediction accuracy. The C4.5 / C5.0 / J48 classifier is among the most popular and powerful decision tree classifiers. C4.5 creates an initial tree using the divide-and-conquer algorithm. The full description of the algorithm can be found in any data mining or machine learning books such as (Han *etal*., 2011) and (Witten *et al*., 2011). C5.0 is an improved version of C4.5 algorithm. WEKA toolkitpackage has its own version known as J48. J48 is an optimized implementation of C4.5 rev. 8.

4. DATA COLLECTION:

The dataset used research was collected from six basic schools in Mafraq city in Jordan. The data are collected from the regular students who are studying in basic schools. The dataset consist of 248 instances. Each instance consists of four attributes; the average grade of the 10th class, the average of the 10th , 9 th , and 8 th classes, and the threshold ratio of the minimum grade acceptable for each track. For example the student to be accepted in science or information track, the average must be equal or greater than 72. Sample of the dataset is presented in Table 1 Where the AVERAGE attribute is the average of the 10th class, the AVG89_10 attribute represents the average of the 8th, 9th , and 10th classes, and the Ratio attribute which represents the minimum grade acceptable for each track. The ratio thresholds are set by the schools administration to distribute students over the track.

Table 1: Sample of the collected Data.

AVERAGE	AVG89_10	Ratio	Branch _accepted
90	88	>=72	Science
81	77	>=72	Management
72	69	>=58	Academic
61	56	>=55	Profession
60	57	>=72	Science
54	50	>=50	Profession

5. THE USED TOOL:

WEKA toolkit (Witten *et al*., 2011) is a widely used toolkit for machine learning and data mining originally developed at the University of Waikato in New Zealand. It contains a large collection of state- of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for regression, classification, clustering, association rules, visualization, and data pre-processing. WEKA has become very popular with academic and industrial researchers, and is also widely used for teaching purposes. To use WEKA, the collected data need to be prepared and converted to (arff) file format to be compatible with the WEKA data mining toolkit.

6. BUILDING THE MODEL:

The classification model is built to give a guideline to help students and school management to choose a suitable track, by analyzing the experience of previous students with similar academic achievements. This model aims to improve the quality of education. The technique of decision tree was applied. Decision tree was developed to determine which track is more suitable for a student. The WEAK toolkit was used to build the decision tree using the C4.5 algorithm (J48 in WEKA), by selecting the best attributes using the information gain measure. The generated decision tree is shown in Figure 1. The decision tree in Figure 1 is easy to be read and understood. The oval represents an attribute decision to follow in the tree, whereas the rectangle represents the suitable track. The classification is to find out the most suitable track for a student based on more than one factor such as the Ratio and the average of student mark in the 10th class (AVERAGE), and the average of the student mark in 8 th , 9th, and 10th classes (AVG89_10).

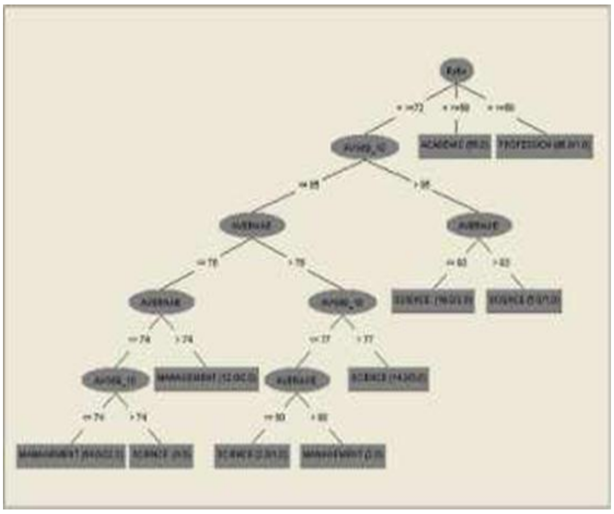


Figure 1: The generated decision tree.

7. RULES EXTRACTION:

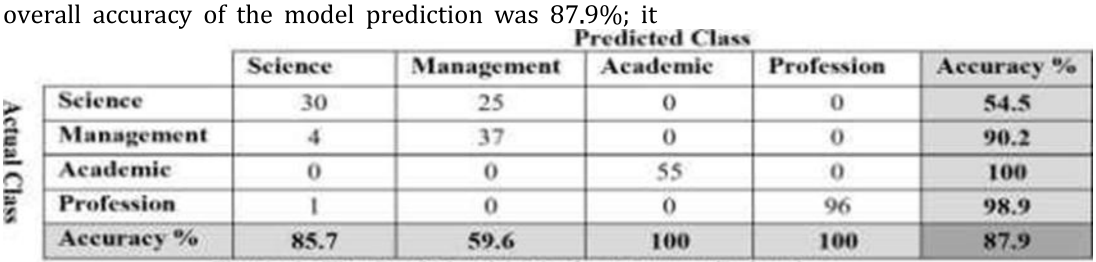
A set of rules can be extracted from the decision tree. These rules are used to predict and classify the class label for each student. The class label in this tree acts as the suitable classified track after the end of a basic education stage. The set of extracted rules from the decision tree are shown in Table 2.

Table 2: The set of classification rules extracted from the decision tree

Rule #	**Rules Antecedent**	**Rule Consequence** **Track =**
1	If *Ratio*>= 50 &< 58	**Profession**
2	If *Ratio*>= 58 &< 72	**Academic**
3	If *Ratio*>= 72 &*Average8,9_10*> 85 &*Average*> 93	**Science**
4	If *Ratio*>= 72 &*Average8,9_10*> 85 &*Average*<= 93	**Science**
5	If *Ratio*>= 72 &*Average8,9_10*<= 85 &*Average*> 78 &*Average8,9_10*> 77	**Science**
6	If *Ratio*>= 72 &*Average8,9 _10*<= 85 &*Average*> 78 &*Average8,9_10*<= 77 &*Average*> 80	**Management**
7	If *Ratio*>= 72 &*Average8,9_10*<= 85 &*Average*> 78 &*Average8,9_10*<= 77 &*Average*<= 80	**Science**
8	If *Ratio*>= 72 &*Average8,9_10*<= 85 &*Average*<= 77 &*Average*> 80	**Management**
9	If *Ratio*>= 72 &*Average8,9_10*<= 85 &*Average*<= 78 &*Average*> 74	**Management**
10	If *Ratio*>= 72 &*Average8,9_10*<= 85 &*Average*<= 78 &*Average*<= 74 &*Average8,9_10* > 74	**Science**
11	If *Ratio*>=72 &*Average8,9_10*<= 85 &*Average*<= 78 &*Average*<= 74 &*Average8,910*<= 74	**Management**

8. MODEL EVALUATION:

To evaluate the generated model, the 10-fold Cross Validation method was used. The generated confusion matrix is presented in Figure 2. We concluded that the overall accuracy of the model prediction was 87.9%; it indicates that the model could correctly classify 218 students among 248 students. It can be noticed that the prediction accuracy of the Academic track is 100%



Actual Class \ Predicted Class	Science	Management	Academic	Profession	Accuracy %
Science	30	25	0	0	54.5
Management	4	37	0	0	90.2
Academic	0	0	55	0	100
Profession	1	0	0	96	98.9
Accuracy %	85.7	59.6	100	100	87.9

Figure 2: The Confusion Matrix for Accuracy Estimation.

9. CONCLUSION:

It can be concluded from the above study the authors proposed and built a simple classification model to provide a guideline to help students and school management to choose the right track of study for a student. This model aims to improve the quality of education where the model is intended to help student to choose the suitable track of their study, by analyzing the experience of previous students with similar academic achievements. For this purpose, the decision tree was used to build the model that is used to determine suitable track for students. This research should be further enhanced as a future work by considering data from several other schools in other cities in Jordan and collect more instances to build the model. Other attributes could also be added to the data set for further enhancing the generated model. Furthermore, some other classification models could be tested in this domain.

REFERENCES:

[1]. Han J., Kamber M., and Pie J. 2011. Data Mining Concepts and Techniques. 3 rd edition, Morgan Kaufmann Publishers.

[2]. Waraporn J. 2009. Classification Model for Selecting Undergraduate Programs, Eighth International Symposium on Natural Language Processing, IEEE

[3]. Tissera R., Athauda I., and Fernando C. 2006.Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining, ICIA, IEEE.

[4]. Nguyen N., Paul J., and Peter H. 2007. A Comparative Analysis of Techniques for Predicting Academic Performance. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12.

[5]. Al-Radaideh Q., Al-Shawakfa E., and AI-Najjar M. 2006. Mining Student Data using Decision Trees, In Proceedings of the International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.

[6]. Cesar V., Javier B., liela S., and Alvaro O. 2009. Recommendation in Higher Education Using Data Mining Techniques, In Proceedings of the Educational Data Mining Conference.

[7]. Pathom P., Anongnart S., and Prasong P. 2008. Comparisons of Classifier Algorithms: Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students, Sripatum University Chonburi Campus, Office of Computer Service, Chonburi Thailand, IEEE.

[8]. Muslihah W., Yuhanim Y., Norshahriah W., Mohd Rizal M., Nor Fatimah A., and Hoo Y. S. 2009. Predicting NDUM Student's Academic Performance Using Data Mining Techniques, In Proceedings of the Second International Conference on Computer and Electrical Engineering, IEEE computer society.

[9]. Naeimeh D., Mohammad S., and Mohammad B. 2004. A New Model for Using Data Mining Technology in Higher Educational Systems, In Proceedings of the IEEE Conference.

[10]. Naeimeh D., Mohammad B., and Somnuk P.2005. Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System, In 11. Naeimeh D., Somnuk P., and Mohammad B. 2008. Data Mining Application in Higher Learning Institutions, Informatics in Education, Vol. 7, No. 1, pp. 31–54.

[11]. Ramaswami M., and Bhaskaran R.2010. CHAID Based Performance Prediction Model in Educational Data Mining, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1.

[12]. Yiming M., Bing L., Ching W., Philip Y., and Shuik L. 2000. Targeting the Right Students Using Data Mining, ACM.

[13]. Fadzilah S. and Abdoulha M. 2009.Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining, In Proceedings of the Third Asia International Conference on Modelling & Simulation Conference, IEEE computer society

[14]. Witten, I., Frank E., and Hall M. 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3 rd Edition, Morgan Kaufmann Publishers.