Proceedings of the Federated Conference on Computer Science and Information Systems pp. 113–117
DOI: 10.15439/2017F283
ISSN 2300-5963 ACSIS, Vol. 11

Measurement of the appropriateness in career selection of the high school students by using data mining algorithms: A case study

Hidayet Takci, Kali Gurkahraman, Ahmet Firat Yelkuvan
Cumhuriyet University
Sivas, Turkey
Email: htakci, kgurkahraman, aftyelkuvan@cumhuriyet.edu.tr

***Abstract*—Less than optimal choice of the university department is one of the serious problems Turkish high school students have been suffering. There are a number of potential factors affecting the student's choice of her future profession. Some of these have received attention in the literature, but such studies do not always involve an investigation of the relationship between the factors analyzed and subsequent levels of academic achievement. The present study examines the relationship between the level of academic achievement and the students' abilities, interests and expectations, by using different data mining methods and classifiers, as a preliminary work to develop a system that will guide the student to selecting a career that will be a better match for her in the future. C4.5, SVM, Naive Bayes and MLP algorithms are used for the analysis; 10-fold cross validation and train-test validation are used as models to evaluate the classifiers results. The student feature set is obtained through questionnaires and psychometric tests. The questionnaire and the psychometric test were applied to 210 and 52 students respectively, from the Computer Engineering Department at Cumhuriyet University. The class was labeled either "successful" or "unsuccessful" with reference to the grades received by each student in computer engineering courses. The comparisons of various data mining algorithms, different data set results, and models used are presented and discussed.**

I. Introduction

IN Turkey, all high school students are subjected to a central multiple choice examination in order to establish whether he or she is sufficient to study in a predetermined field. The examination is organized by an institution called in Turkish language "Ogrenci Secme ve Yerlestirme Merkezi" (OSYM). OSYM prepares a guide to introduce the departments and universities to the candidates. The candidates specify their choices for the suitable departments by examining the information in this guide as well as other factors.

Although there is no information in the OSYM guide, the important factors for choosing a department are student ability, interest, and demographic data. For example, there is a close relation between mechanical ability and some engineering fields. The abilities of the candidate such as abstract thinking, understanding the shape relations, mechanical skills, hand-eye coordination and creativity should be measured properly. Determining the student interests is also important since considering only the ability of the student to find out the proper profession may not be enough. Therefore, searching for matches between two sets, one of which include abilities, interests, and demographic data and the other one includes student success, should be done. The interests of a student may be sciences, social sciences, agriculture, and business trade. A suitable situation for the student is to have a profession in which he or she is interested as well.

Relatively more studied subject in the literature is student expectations from the profession [1,2]. The prominent factors of the student expectations are for example allowance of his or her family, social benefit, social expectations, career opportunities, and salary. Once profession choice is made, there may be many options of university in which the student will study. For this reason, possibilities and sufficiency of the university department should meet the student expectations. The quality and quantity of teaching staff, whether the university has a student exchange agreements, scholarships, success level are examples of factors that a student takes into consideration in order to choose the suitable university. Another criterion is the expectations of the department from the student. This is an important subject for student success in this department. A profession field specifically requires a competence to such as math, social or art. Other possible expectations from the students are ability to cope with stress, sex, health status, educational level of the family, type of high school, grade point average in high school etc.

In Turkey, in most of the high schools, the guidance counselors or teachers practically only examine the student grades in different lessons to determine suitable departments that the student can choose. The lack of professional guidance system which determines possible proper carriers by considering all the parameters related to student may cause low academic performance or dropouts in the future.

In this study, preliminary study has been done in order to build a system that can reveal the relationship between student and department features. For this purpose, educational data mining (EDM) algorithms have been run on data collected from students and departments. EDM can help to take into account many parameters, recognize the most important ones and figure out their relationship in order to understand and

solve educational questions [3,4]. Input data of our model are the information of students, and the output is basically whether this choice is suitable or not.

II. Related Works

Data Mining (DM) is a field that finds out new and useful information from a high volume data [5]. Having many application areas such as marketing, medicine and real estate etc. it has also suitable techniques for educational practices as well. Many data mining techniques are applicable to educational fields and it is specifically called EDM. Shu-Hsien et al. [3] reviewed the EDM studies of period between 2000 and 2011 in 9 different categories while Romero and Ventura [4] reviewed the studies of period between 1995 and 2010 as DM and Computer Based Educational System (CBES) categories.

In addition to DM techniques, statistical and machine-learning algorithms are also used in EDM in order to process the educational information for two main objectives which are improving student performance and educational environment by studying students' approaches and examining the educational methods. For these aims, EDM makes some evaluations and predictions by using many types of data from different sources such as student information system which may include student grades and other personal information, questionnaires about learning process and lessons, tutoring systems, on-line educational web applications and some educational software that students use, suck as for following the lecture notes and homework. For instance, determination of common lessons taken by students, whether a student can pass a specific lesson or not and classification of students can be made by using dense pattern mining [6], classifiers [7] and prediction models [8] respectively. Apart from the classical data mining applications, psychometric properties of students are also used in educational implementations [9].

DM has been dealing on student performance more than other subjects since it is an important and popular topic in education. Recently, various studies have also been made to improve the quality of lecture books [10], to increase the student attendance to lessons by using social network analysis [11] and to integrate the students' information in educational software [12]. For a more detailed example, Maria et al. [13] studied on log files of a free web-based tutoring system for middle school mathematics which was being used by 3,747 students in New England to predict student attendance to college by using logistic regression. They reported that their system can distinguish a student who will enroll in college 68.60% of the time. In India, Yadav and Pal [14] applied C4.5, ID3, and CART decision tree algorithms on engineering student's data to predict whether a student will pass, fail or promote to next year. Accurate classification rate of the study reached 67.78% for C4.5. Quadri and Kalyankar [15] studied on student dropouts features according to student risk factors such as gender, attendance, previous semester grade and parent income etc. Their hybrid method uses combination of the decision tree algorithms and logistic regression.

The studies in EDM field have concentrated on student performance and how to enhance learning process. Although these topics are very important for educational life, one of the main failure causes is the student incorrect department or career choice. In this study, we proposed a model using EDM techniques to help high school students in their career choices by considering different student information including student interests, abilities, student, and department expectations, demographic and, psychometric test data.

III. Predection Model

The test data and the output of the model are temporarily limited to our student still studying in Computer Engineering Department in Cumhuriyet University. The information of students and departments were collected from questionnaire and psychometric tests applied to our students, student information system in our computer engineering department. All the raw data should be preprocessed in order to eliminate redundancy and to convert them in suitable formats for processing by data mining techniques. Pre-processing stage also includes extracting features of students.

Processed data are analyzed by using data mining algorithms to obtain rules and patterns. C4.5 and regression analysis are suitable algorithms within rule-based and score-based classifiers respectively. The outputs of data mining stage are rules and patterns. In assessment stage, these rules and patterns are reviewed to eliminate the weak rules. Obtained valuable rules and patterns are used in EDM process.

The characteristic properties of students such as demographic information, their psychometric features, and information obtained from the guidance department and the factors affecting the department choice are the predictor variables for the model. The output variable is about relevance or achievement of the student for the department.

Prediction models have been used for testing the proposed system. In this study, historical data consisting of student abilities, interests, demographic information and student expectations from the department were obtained by questionnaires and psychometric tests applied to the students. Psychometric tests were basically used in order to be able to find out the student abilities while questionnaires included questions to provide the entire student related features. The achievement of the student in the department was used as the class label. The class label was "successful" or "unsuccessful" according to the computer related lecture grades of each student.

Two important components of the system which are feature set for matching and machine learning algorithms used in the system are presented.

A. Feature Set

We have features for both questionnaire and psychometric test. While feature selection was made for 115 questions of questionnaire, no feature selection was needed in psychometric test since there were already 20 features in the test.Backward-logit, forward-logit, fisher filtering and reliefF methods have been used for feature selection.

1) Questionnaire feature set: Typeset sub-subheadings in medium face italic and capitalize the first letter of the first word only. There are totally 115 questions in the questionnaires and the categorization is as the following according to their measurement objective.

- 16 questions for student abilities
- 14 questions for student expectations
- 10 questions for demographic information
- 75 questions for student interests

The titles for category of measuring student abilities are as follows.

• Hand-eye coordination • Visual-spatial relation
• Logical-mathematical • Verbal-linguistic

Student expectations are related to factors affecting the profession such as career, salary, job opportunities, flexible working etc. The category of measuring student interests which contains most of the questions is considered as having following interest titles.

• Verbal-linguistic • Logical-mathematical
• Social sciences • Agriculture
• Foreign languages • Art • Music
• Literature • Sciences • Business and trade etc.

Although some of interest fields are not related to this study, the aim of DM analysis is to find out the relationship between the factors that are not easily noticeable. Although some of interest fields do not seem to be related to this study, the aim of DM analysis is to find out the relationship between the factors that are not easily noticeable. Therefore, the analysis was performed with interest titles mentioned above in the study.

In Table 1, grade column indicates the grading method of the questionnaires. Linear scale method is used in our questionnaires. In both talent related perception questionnaire and interest ares questionnaire 1 means "always", 2 means "often", 3 means "sometimes", 4 means "rarely" and 5 means "never". Also in expectations questionnaire 1 means "very high", 2 means "high", 3 means "average", 4 means "low" and 5 means "very low".

2) Psychometric data feature set: The quality of questionnaire-based measurement is no doubt dependent on the student answers. Therefore, psychometric test which has answers with naturally less indiscrimination was decided to perform in order to compare its analysis results with the ones obtained by questionnaire. Psychometric test with 20 questions aims to reveal student abilities such as logical-mathematical, comprehending visual-spatial relation, eye-hand coordination, and memory usage. The skills and the number of related questions are as follows.

• Logical-mathematical(4) • Verbal-linguistic(2)
• Imagination(2) • Memory usage(3)
• Visual-spatial relation(4) • Attention(3) • Mechanical(2)

For example, a student is asked to look at a village image for a while. Then some questions asked to the student such as "What is the number of houses?" and "What is the color of the bridge?" in the picture. With these questions, the memory usage and attention ability of the student is measured.

TABLE I
EXAMPLES QUESTIONS OF QUESTIONNAIRES

Questionnaire	Questions	Grades				
Talent Related Perception Questionnaire	I can do simple mathematical operations in the mind.	1	2	3	4	5
	I try to learn the meanings of the words I just heard.	1	2	3	4	5
Expectations Questionnaire	How important is a peaceful business environment to you?	1	2	3	4	5
	How important is economic prosperity for a profession to you?	1	2	3	4	5
Interest Areas Questionnaire	Do you enjoy reading novels in history?	1	2	3	4	5
	Would you like to take apart a tool and reassemble it?	1	2	3	4	5
	Would you like to try growing new flower species?	1	2	3	4	5

B. Predictive Data Mining Algorithms

Different DM models can be applied to student-department matching. In this study, prediction models were used. One of them that was used is called C4.5 and it is based on decision tree analysis. Statistical-based Naive Bayes algorithm and a successful classifier called Support Vector Machine (SVM) are the other models that were used in this study.

C4.5 is a classifier based on Quinlan ID3 algorithm and usually used in classification studies [16]. It constructs decision trees from a set of labeled data by using information gain. C4.5 is usually used in EDM since it is easily comprehensible and its results are relatively simple to conclude [17,18]. Naive Bayes classifier approaches the problem in probabilistic manner. While it is generally used in pattern recognition, it has been using also in EDM applications [17,19]. It considers each feature as independent of the others. It is also known to be fast algorithm and have high accuracy results.

SVM was first introduced by Vapnik and et al. [20] and it has been used in many classification applications in view of obtaining high accuracy results. In SVM, linear separability is a manner issue. First, the input space is mapped to a kernel space. Then, kernel space is used to constitute a linear space. In SVM based classification model, the classes are separated as the farthest possible points from each other.

IV. EXPERIMENT DESIGN

In this section, the details of dataset and feature sets, a brief explanation of the experiment design, and the experiment results are presented.

A. Data Set

We have two data sources which are from questionnaires and psychometric test results. The questionnaire and psychometric test were applied to 210 and 52 students from Computer Engineering Department in Cumhuriyet University. There were totally 210 students under questionnaire but 4 of them were lost data. The 52 students who were tested psycho-metrically are a subset of the surveyed students. The reason for this small number of elements in the lower cluster is that the test takes a long time.

B. Experiment Design and Results

Using three different DM algorithms mentioned above, the results were obtained by analyzing both questionnaire data consisting of 5 different categories and psychometric basic feature dataset for each algorithm.

TABLE II
THE CLASSIFIERS' ACCURACY RESULTS OF QUESTIONNAIRE DATASET CONSISTING OF 5 DIFFERENT CATEGORIES

Variables	Validation Method	C4.5	SVM	Naive Bayes	MLP
All Variables	10-fold validation	59.50	62.00	61.50	59.00
	Train-test validation	61.29	54.84	67.74	70.97
Ability Based	10-fold validation	55.00	69.50	64.50	63.00
	Train-test validation	48.39	67.74	69.35	62.90
Student expectations based	10-fold validation	55.00	69.00	70.50	68.50
	Train-test validation	64.52	72.58	67.74	72.58
Demographic data based	10-fold validation	62.50	-	63.00	-
	Train-test validation	69.35	-	69.35	-
Student interest based	10-fold validation	59.00	59.50	62.00	57.50
	Train-test validation	50.00	64.52	59.68	56.45

TABLE III
ACCURACY RESULTS FOR QUESTIONNAIRE DATASET

	10-fold cross validation	Train-test validation
Algoritm	*Accuracy(%)*	*Accuracy(%)*
C4.5	59.50	61.29
SVM	62.00	54.84
Naive Bayes	61.50	67.74
MLP	59.00	70.97

For each feature set, accuracy ratios are given and ROC analyses are performed. In the experiments C4.5, Support Vector Machine (SVM), Naive Bayes, and Multilayer Perception (MLP) algorithms were used. 10-fold cross validation and train-test validation were used for model evaluation. 90% of the data set was used for training, and the remaining 10% was used for test. 10-fold cross validation is repeated ten times and averaged.

Since accuracy rate may not be competent in some cases, ROC analysis was also performed on both questionnaire and psychometric data. Thus, the classifier producing the best results for each data set was found out.

1) Analysis of Questionnaire Dataset: In this study, for the comparison of algorithms, the variables in the questionnaire dataset were analyzed in 5 categories according to their bases which are ability, student expectation, demographic data, student interest and all variables including all the bases. The analysis results are shown in Table 2.

However, since some of the demographic data such as residence region of the student's family and high school that the student graduated, are not numerical or categorical variables, were not used in SVM and MLP classifiers. For this reason, results were obtained only from C4.5 and Naive Bayes algorithms for demographic data. In addition, demographic data are not also used in questionnaire dataset consisted of all variables because of the same problem. MLP algorithm achieved the highest accuracy values according to train-test validation according to the results in Table 2.

As a first step, performances of machine learning algorithms were measured according to questionnaire data based on all variables. The accuracy value was preferred as performance criteria for the algorithms. 10-fold cross validation and train-test validation were used to evaluate the classifiers results. The analysis results can be seen in Table 3.

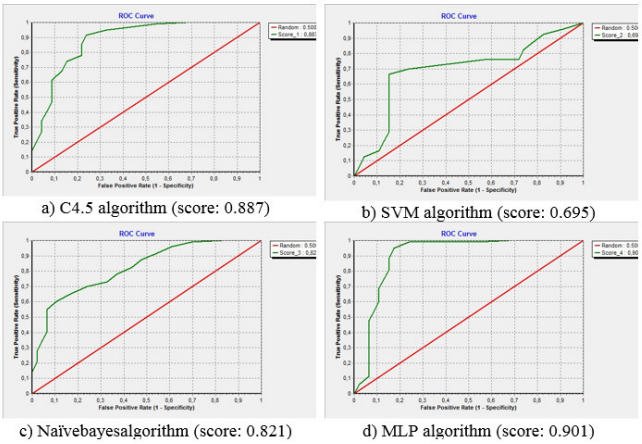

Fig. 1. ROC analysis results for the algorithms applied to questionnaire dataset

TABLE IV
THE CLASSIFIERS' RESULTS OF PSYCHOMETRIC DATASET

	10-fold cross validation	Train-test validation
Algoritm	*Accuracy(%)*	*Accuracy(%)*
C4.5	54.00	62.50
SVM	56.00	68.75
Naive Bayes	56.00	56.25
MLP	48.00	50.00

According to the results for 10-fold cross validation, although SVM which produced the highest accuracy value as 62.00%, it is not possible to describe it as the best classifier since the values of all algorithms are close to each other. In train-test validation, the result of MLP is significantly superior to other algorithms. In 10-fold cross-validation, a further ROC analysis was needed because close values were seen. The ROC curves obtained for each algorithm are presented in Figure 1.

The method used to evaluate ROC curves is to find the area under the curve (AUC). In Figure 1, score values refer to these AUC values. As in the train-test validation, according to the score values, MLP algorithm gave the best classification result. Therefore, MLP can be accepted as the most suitable algorithm for the questionnaire dataset.

2) Analysis of Pyscometric Dataset: Both the accuracy and ROC analysis were also performed for psychometric test dataset. According to the accuracy ratios, both questionnaire and psychometric test results are similar. The classifiers' results for psychometric test dataset can be seen in Table 4.

SVM algorithm gave the best accuracy value as 68.75% according to train-test validation in Table 3. However, SVM algorithm could not perform high result in 10-fold cross validation. Therefore, ROC analyses were also used to measure classifiers' performance.

As can be seen in Figure 2, although SVM has the highest accuracy ratio, AUC value of C4.5 algorithm is the best value as 0.823 according to ROC curves. This result is also different to questionnaire dataset analysis which selected MLP algorithm.

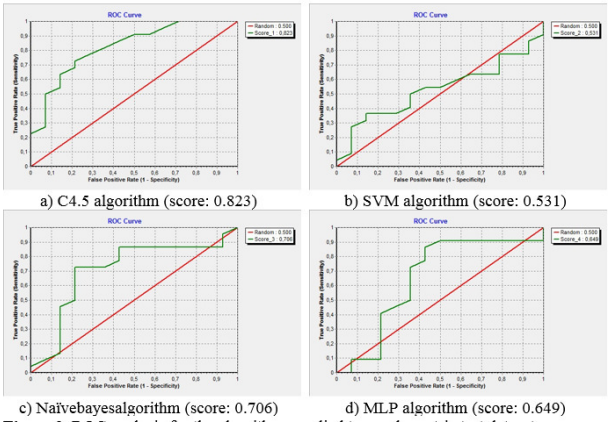

Fig. 2. ROC analysis for the algorithms applied to psychometric test dataset

V. Conclusion

The analysis results of the questionnaire data were better than the ones obtained from psychometric data. The main reason for the superiority of the analysis results is that the questionnaire data was obtained from the many aspects of students such as abilities, interests, demographic data, and expectations. Among questionnaire data, demographic data gave more discriminative information than the others. Besides, the relationship between ability and success appeared to be more valuable than the relationship between interest and success. Although SVM generally has mentioned as it is superior in literature, C4.5 and MLP algorithms gave better results in this study.

According to model comparison, the success obtained by Train-Test method is quite better than the result obtained by 10 fold cross-validation method.

In the literature on the educational studies, as far as we can know, there is no study investigating the relationship between academic achievement and the topics covered in this study by using EDM algorithms. On the other hand, there are studies that are relevant in terms of their contents. The main axis of the studies is what the students take into consideration when choosing a university department and a career or what are the factors that influence them [1,2]. The studies in this area have generally examined the variables that influence the choice of the students and attempted to find out which factor is more discriminative. In both studies, in order to reveal the effective factor, the students' interests and expectations as well as other factors are scaled in the questionnaire. Misran et al. also includes student demographic information in the study. The difference of our study from the others is to focus on the proper choice of the student which otherwise could lead to low academic performance in the future.

Acknowledgment

We would like to thank Turkish Scientific and Technological Research Council (TUBITAK) for providing the research support (Project Number: 115E837).

References

[1] N. Misran, N. Abd.Aziz, N. Arsad, N. Hussain, W. Zaki and S. Sahuri, "Influencing Factors for Matriculation Students in Selecting University and Program of Study.", Procedia-Social and Behavioural Science, vol. 60, pp. 567-574, 2012.

[2] C. BobAlca, O. Tugulea and C. Bradu, "How are the students selecting their bachelor specialization? A qualitative approach.", Procedia Economics and Finance, vol. 15, pp. 894-902, 2014.

[3] L. Shu-Hsien, C. Pei-Hui and H. Pei-Yuan, "Data mining techniques and applications - a decade review from 2000 to 2011.", Expert Systems with Applications, vol. 39, pp. 11303-11311, 2012.

[4] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art.", IEEE Transactions on systems, man, and cybernetics, part C: applications and reviews, vol. 40, pp. 601-618, 2010.

[5] I. Witten and E. Frank, Practical Machine Learning Tools and Techniques with Java Implementations., 1st ed. San Francisco, CA, USA: Morgan Kaufmann, 1999.

[6] O. Zaine, "Web usage mining for a better web-based learning environment.", in Conference on Advanced Technology for Education, Banff, Alberta, Canada, 2001, pp. 60-64.

[7] H. Cha, Y. Kim, S. Park, T. Yoon, Y. Jung and J. Lee, "Learning styles diagnosis based on user interface behaviours for the customization of learning interfaces in an intelligent tutoring system", in 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 2006, pp. 513-524.

[8] W. Hamalainen and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems.", in 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 2006, pp. 525-534.

[9] P. Pavlik, H. Cen, L. Wun and K. Koedigner, "Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor.", in 1st International Conference on Educational Data Mining, Montreal, Quebec, Canada, 2008, pp. 77-86.

[10] R. Agrawal, S. Gollapudi, A. Kannan and K. Kenthapadi, "Data mining for improving textbooks.", ACM SIGKDD Explorations Newsletter, vol. 13, pp. 7-19, 2011.

[11] [3]R. Rabbany, M. Takaffoli and O. Zaiane, "Social Network Analysis and Mining to Support the Assessment of On-line Student Participation.", ACM SIGKDD Explorations Newsletter, vol. 13, pp. 20-29, 2011.

[12] Z. Pardos, S. Gowda, R. Baker and N. Heffernan, "The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software.", ACM SIGKDD Explorations Newsletter, vol. 13, pp. 37-44, 2011.

[13] M. San Pedro, R. Baker, A. Bowers and N. Heffernan, "Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School.", in 6th International Conference on Educational Data Mining;, Memphis, TN., USA, 2013, pp. 177-184.

[14] S. Yadav, S. Pal, "A Prediction for Performance Improvement of Engineering Students using Classification.", World of Computer Science and Information Technology Journal , vol. 2, pp. 51-56, 2012.

[15] M. Quadri, N. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques.",Global Journal of Computer Science and Technology , vol. 10, pp. 2, 2010.

[16] J. Quinlan, C4.5: Programs for Machine Learning, 1st ed. San Francisco, CA, USA: Morgan Kaufmann, 1992.

[17] Q. Al-Radaideh, E. Al-Shawakfa, M. Al-Najjar, "Mining student data using decision trees.", in International Arab Conference on Information Technology ;,Yarmouk University, Jordan, 2006.

[18] S. Yadav, B. Bharadwaj S. Pal, "Data Mining Applications: A comparative study for predicting students' performance.",International Journal of Innovative Technology and Creative Engineering , vol. 12, pp. 13-19, 2011.

[19] B. Bharadwaj S. Pal, "Data Mining: A prediction for performance improvement using classification.",International Journal of Computer Science and Information Security, vol. 9, pp. 136-140, 2011.

[20] C. Cortes, V. Vapnik, "Support-vector networks. "Machine Learning , vol. 20, pp. 273-297, 1995.