Volume 4 No. 1, January 2014 ISSN 2223-4985

International Journal of Information and Communication Technology Research

http://www.esjournals.org

Prediction the Loyal Student Using Decision Tree Algorithms

[1]Saeide kakavand, [2]Taha Mokfi, [3]Mohammad Jafar Tarokh
[1,2]Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
[3]Department of Industrial Engineering, K.N. Toosi University of Technology, Tehran, Iran

ABSTRACT

One of the most important challenges that higher education system facing today is Providing more effective, efficient and higher quality education service to students, and predicting the pattern of loyal students. Because the universities are trying to raise educational quality, Applying data mining in higher education helps the manager, lecturer, and students to make higher performance. The aim of research paper is to understand the external factors that may cause the student loyalty. By doing that, the university can identify students who have decided to continue studying, so it can invest on them, and thus increase its educational quality. One of the best ways to achieve this is by using valid management and processing of the students database.

In this study, using dataset from the Private University and applying data mining techniques, classify master students based on input characteristics and finally the pattern of faithful students (students who have decided to continue studying)were extracted. Classified students are based on personal information of students, student academic status, type of their pervious university (private or state university), finances and occupation status, and educational status of their parents. To classify students, the rule generation process is based on the decision tree algorithms like C.5, CART and CHAID. The results showed that CART decision tree algorithm is the best predictor with 94% accuracy on evaluation sample.

Keywords: *component; data mining, decision tree algorithms, loyal student, higher education, student's performance*

1. INTRODUCTION

Nowadays, the higher education systems and the amount of educational data that used in decision making Processes have been evolved. Therefore, educational organizations start developing and improving the educational systems .The best decisions can be made by using the new techniques such as data mining methods. Data mining is the process of extracting useful knowledge from amount of data that are collected in databases. Considering that in the majority of universities prepares a massive database of student's specifications that can include information and valuable models. This information contains students and their family characteristics, educational and academic backgrounds. Studying the hidden patterns and knowledge of this information can be helpful for decision makers in the higher educational systems to enhance and improve the educational processes in different fields such as scheduling, registering, evaluation and counseling.

Data mining findings about large educational databases are understandable, useful, previously unknown, valid, and innovative. The process of student's registration in any educational systems or discovering the factors that lead to success of students are great concerns for the higher education managers, therefore advanced data mining techniques such as clustering or classification can be used in finding valuable and specific patterns of loyal and successful students.

As a result, researchers try to determine the variables that are related to academic achievement of students and may affect the registration process. Therefore, one of the most important challenges that higher education faces is recognizing the pattern of loyal students (Students who have decided to continue studying in PhD level).

This research paper is an effort to use the data mining processes, particularly tree decision algorithms to determine pattern of loyal students and therefore improve the quality of the higher educational system. The main goal of this paper is to provide accurate, practical and reliable models that can meet the fundamental requirements of universities. It is thought that the factors that affect achievement could be useful for students, instructors, and administrators who are interested in achieving success. This paper is presented as follows; in the coming section (Section 2) a literature review of applying data mining in higher education is explained. In Section 3 the research methodology is given, where data, modeling, and analysis are explained in detail. In Section 4, the comparative analyses of the models and the results of model analyses are presented, and finally, in the last section (Section 5) the conclusion is given.

2. LITERATURE REVIEW

Data mining is applied in various researches in different fields of education. Determining the factors that affect students' academic achievement is an essential input to improve the educational systems, thus a great deal of the prior studies analyzed this case [1], [2]. They tried to collect data, usually from survey type tools, to figure out the correlation between factors and their impact on academic achievement. In addition, researching on students retention has been survey driven (e.g., surveying a group of students and determine whether they continue their education or not). For instance some researchers focused on gifted education [3] some others studied the correlation between academic achievements and parenting styles [4], others investigated Forecasting Enrolments and Student Retention [5],…,[13] One of the first papers that used data mining to predict students' enrolment has been written by Song, et al in 1993.They Forecasted Enrolment students with Fuzzy Time Series. For evaluating the forecasting model, they used ordinary linear regression method, and predicted values obtained from fuzzy set compared with the actual results. Jing Luan could predict and cluster loyal (successful) and unsuccessful students by using two-step clustering algorithm, decision tree algorithms and neural networks on data from 15000 students. He identified the factors affect students retention and academic performance. He used data mining to find the profile pattern of unsuccessful students. He found the main attributes that may be associated with dropouts by using feature selection and association rules, also he identified potential "at risk" students by utilizing Classification and clustering.

Mostly researchers attempt to predict performance of a student [14], [15], [16]. For instance, Menzel and Bekele used Bayesian networks for predicting performance of a student, based on some identified attributes. This research specifically focused on personal, social and cultural features that may be used in automatic prediction of performance. Elements that are involved in the educational objectives include: teaching strategies/learning, parents, teachers and students.

In addition, Superby and Vendome found out the factors influence the achievement of the first-year university students by using Data Mining Methods [17]. They provided the most significant variables related to academic success among the entire questionnaire that is asked on 533 freshmen. Finally, they presented the results of the application of discriminant analysis, neural networks, random forests and decision trees aimed at predicting those students' academic success. Some researchers claimed to find strong correlation between student's success and family income. Their results indicated that the amount of income has a positive effect on success and academic achievements [18], [19]. They indicate that children who are born to well-educated parents have more opportunities than those born to poorly-educated parents.

Some other researchers investigated the effect of teacher support upon academic achievement [20] while others focused on the importance of different schools types [21]. Moreover, other researchers used data mining in higher educational systems [22], [23].

Therefore, with the help of experts and according to major studies in the fields of education is about predicting Student performance and the final score or identifying student enrolment patterns using data mining techniques.

However these papers have used data from the first year undergraduate students. Due to this, in this study, we tried to apply data mining techniques such as decision trees algorithms on master student's data, like family characteristics, educational and academic backgrounds, to identify the pattern of loyal students (students who are planning to study at the PhD level).

3. MATERIALS AND METHODS

In this research study, we followed a popular and best known data mining methodology called CRISP-DM (Cross Industry Standard Process for Data Mining)[24].

A. Data

Data is the principal subject of a knowledge discovery. The selection of attributes and their types have a strong effect on model accuracy. The data used in this study was collected from Master students. The attributes were general information about the students such as demographic characteristics (gender, age, marital status, registered city, etc.), employment status, educational background (type of previous university, average, the number of terms spent at previous university, etc.), parents education, and financial. The data extracted contained 14 attributes with demographical and academic information of the students. A complete list of attributes obtained is given in "Table 1".

Table I: Student Related Variable

Field	value
Gender	Male, Female
Age	20-37
marital status	Single , married
Registered City	Local , outsider
Previous university	State, private
Average	12 – 19.08
the number of terms	7 - 14
Mother's education	low literacy, Diploma, Associate degree, Bachelor degree, Master, PhD
Father's education	low literacy, Diploma, Associate degree, Bachelor degree, Master, PhD
Employment state	Jobless, employee , self employed
Amount of income	0 – 20 million RLS
Continue studying at PhD level	Yes, No

I. Pre-proccesing data

Data preparation process included a comprehensive analysis of the variable and their values in order to remove or reduce distorted data (e.g., missing, noisy, outliers or incomplete values. Sometimes some attribute values are missing, so there are two basic solutions: we can either remove the whole attributes that had a large number of missing value or replace the missing value with a constant, or mean value. In this research, because some of the income attribute values were missing, so we had replaced Average earnings for every different group of job (Jobless, employee, self-employed). After completing data cleaning, the attributes are converted to appropriate representational form for data mining algorithms. In the data transformation phase, the output variables (i.e., age and average) are also transformed into a four-level categorical variable.

II. EDA(Exploratory Data Analysis)

In educational domain, the feature selection is especially important [25]. It is a method to identify relevant and most important features from the full set of attributes. The irrelevant attributes can have impacts on the prediction accuracy [26].

In this study, the important features are ranked based on Pearson Chi-Square. It measured independence of the target and the predictor without indicating the strength of any existing relationship;" Table.2" presents the most important and relevant attributes. Furthermore, the relationship between the mother education and students performance is indicated in *"Fig. 1"*.

Table II: The Most Important Features

Field	Value
Previous university	1.0
Mother Education	1.0
Average	0.992
Father Education	0.974
the number of terms	0.962
Gender	0.941
Age	0.76
Registered City	0.71
Employment state	0.625
Amount of income	0.354
marital status	0.12

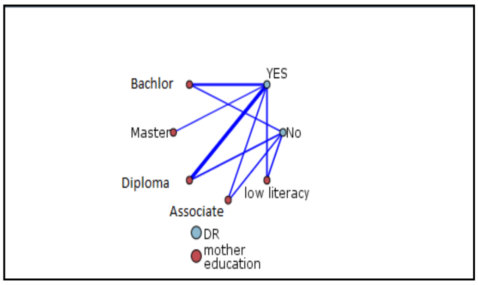


Figure I: Relationship between Mother Education and Students' Performance

B. Select the most proper Classification techniques

The most proper classification technique must be chosen in the first step of modeling process. Classification is an example of the more general problem of pattern recognition. In other words, classification consists of examining the characteristic of a new object and assigning it to one of the predefined classes [27]. In the student classification process, each student will be assigned to a pre-defined group (loyal/disloyal) according to his/her characteristics. In this study, students are classified according to their characteristics and three popular decision tree algorithms are preferred as they provide practical help for more intelligible results (and compared to each other): CART, C.5 and CHAID. These prediction methods are selected because among other prediction methods, decision trees have several considerable advantages: The obtained models are simple and easy to understand [28], and the processes can be performed even with a little effort from users for data preparation [29]. Therefore, decision trees can easily be merged with the information technology. What follows is a concise explanation of the decision tree and the algorithms which are used in this study:

The decision tree learning is a popular method used in data mining, because the structure of decision tree classifiers does not need any domain knowledge or parameter setting; therefore, it is suitable for investigative knowledge discovery [26]. The goal is to generate a model that predicts the value of a target variable based on several input variables. In other words, decision trees are regarded as easily understandable models because there is a reasoning process for each conclusion. A decision tree can be directly converted into a set of IF-THEN rules which are one of the popular forms of knowledge representation. Thus, C4.5 and CART algorithms are not complicated to understand and interpret [30].Each algorithm uses an attribute selection measure to select the attribute tested for each non-leaf node in the tree. In general, decision tree classifiers have good precision.

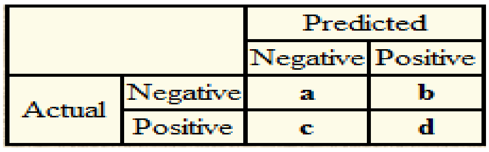
I. Decision Tree Algorithms

C4.5 is an algorithm which is used to build a decision tree developed by Ross Quinlan. The research shows that The C4.5 tree-induction algorithm makes good classification accuracy and is the fastest among the other algorithms [31]. In this study, the C5 algorithm is applied, which is an improved version of C4.5.
Classification and Regression Trees (CART) is a nonparametric method and does not require variables to be chosen in advance [32]. It can easily handle both nominal and continuous attributes as targets and predictors.

The CHAID algorithm is one of the oldest tree classification methods originally introduced by Kass (1980). CHAID (Chi-squared Automatic Interaction Detector) performs multi-level splits when computing classification trees .CHAID can be used for prediction as well as classification, and for detection of interaction between variables. Both CHAID and CART algorithms can be applied to analyze regression-type problems or classification-type.

C. Classifier Accuracy Measures

As much as predictive accuracy is concerned, it is difficult to get general suggestions. As a practical matter, it can be a good idea to use different algorithms and decide on the most reasonable and performing model based on the prediction errors. For comparing the predictive accuracy of models, we applied a popular performance criterion: the prediction accuracy while the percentage of the test set is correctly classified by the classifier. [33], [34] But accuracy is not a reliable metric for the real performance of a classifier, because it produces deceptive results if the data set is unbalanced. So, another way to examine the performance of classifiers is to use a confusion matrix. A confusion matrix for two classes is shown in “*Fig. 2*”. [35].



		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Figure 2: Confusion Matrix

4. RESULT AND DISCUSSION

The prediction results of the three decision tree modeling methods are demonstrated in “Table.3”. Because the target variable had two nominal values, the confusion matrixes show 2*2 square matrix where the correct predictions are put at the diagonal from upper left to lower right corner. The overall accuracy of models is placed at the bottom of the right columns [36], [37].

Table III: Confusion Matrix of Prediction Results for all Classification Models

C.5 algorithm			
Training Partition	**YES**	**NO**	**Accuracy**
YES	67	14	82.71%
NO	4	50	92.59%
		overall	**86.67%**
Test Partition	YES	NO	
YES	19	3	86.36
NO	1	10	90.9
		overall	**87.87**
Validation Partition	**YES**	**NO**	
YES	14	2	87.5
NO	2	17	89.47
		overall	**88.57**
CART algorithm			
Training Partition	**YES**	**NO**	
YES	71	10	87.65
NO	2	52	96.29
		overall	***91.11***
Test Partition	**YES**	**NO**	
YES	20	2	90.9
NO	1	10	90.9
		overall	***90.9***
Validation Partition	**YES**	**NO**	
YES	14	1	93.33
NO	2	18	0.9
		overall	**91.42**
CHAID algorithm			
Training Partition	**YES**	**NO**	
YES	60	14	81.08
NO	4	61	93.84
		Overall	**87.05**
Test Partition	**YES**	**NO**	
YES	26	4	86.66
NO	1	12	92.30
		Overall	**88.37**
Validation Partition	**YES**	**NO**	
YES	10	4	71.42
NO	0	7	100
		Overall	**80.95**

As results indicate, among the three model types, CART decision tree algorithm yielded the best prediction results with 91.42 overall accuracy on the validation set. The CART decision tree is followed by C.5 decision tree with an overall prediction accuracy of 88.57%, Out of the three model types used; the CHAID algorithm produced the lowest prediction accuracy with an overall value of 80.95%.

In addition to estimating the prediction accuracy for each model, the next step is to perform a sensitivity analysis, in order to recognize the relative importance of the independent variables. Each model type produced different sensitivity rankings of the independent variables.

All three sets of sensitivity numbers were collected and then the sensitivity numbers are aggregated into the single “Table.4”. The sensitivity analysis results implied that the most important predictor variables are related into educational background (Previous University, number of the terms) and parent’s educations. As you see, the mother education has noticeable effects on the students’ performance. Some of the demographic attributes like, employment, age, and marital status are not as important as the other independent variables.

Table IV: Aggregated Sensitivity Analysis Results

Attribute	**C.5**	**CART**	**CHAID**	**The averaged sensitivity**
Previous university	0.464	0.535	0.48	0.493
The number of terms	0.113	0.000	0.155	0.089
Gender	0.1	0.018	0.097	0.0716
Mother education	0.129	0.144	0.213	0.162
marital status	0.000	0.029	0.000	0.0096
Registered City	0.000	0.029	0.000	0.0096
Amount of income	0.000	0.029	0.000	0.0096
occupation	0.000	0.029	0.000	0.0096
Average	0.04	0.000	0.000	0.013
Age	0.000	0.111	0.055	0.055

father education	0.153	0.075	0.000	0.076

4. CONCLUSION

One of the most important challenges that higher education faces today is recognizing the paths of students. Data mining is a powerful new technology for a wide variety of real-world problems and industries where amounts of data have been collected. Educational institutes have not utilized it as much as some other fields (e.g., banking, marketing). [38], [39], According to data mining in educational systems, which is an entirely new field of research, it's necessary to develop this methodology for a variety of educational purposes. Some of the noteworthy emerging data mining applications areas in the field of education are gifted student identification, retention management, and improvement of students' performance. As this study demonstrated, data mining techniques can accurately predict loyal (success) students, and therefore its findings allow researchers for doing further analyses and identification in important predictors. Our results indicate that, data mining methods are capable of predicting loyal student approximately 90% accuracy through sufficient given data with the proper features. Among the three individual prediction model which are used, CART decision tree algorithm performed the best followed by C.5 decision tree. Out of the three model types used, the CHAID algorithm produced the lowest prediction accuracy.

Acknowledgment

First and foremost, the author offers her sincerest gratitude to Dr.Frshid Abdi, who has supported her throughout doing this research with his patience and knowledge. And, it is a pleasure to express her gratitude wholeheartedly to the reviewers for their constructive comments.

REFERENCES

[1] Veitch, W. R. "Identifying characteristics of high school dropouts: Data mining with a decision tree model",2004.

[2] Yenilmez, K., Duman, A. " Interviewing with students about the factors that affect the achievement of mathematic in Primary School", Sosyal Bilimler Dergisi,vol 19, pp. 251–268,2008.

[3] Ma, Y., Liu, B., Wong, C.K., Yu, P.S., & Lee, S.M. "Targeting the Right Students Using Data Mining". In Proceedings of the Knowledge and Data Discovery (KDD2000), Boston, USA. pp. 457-464,2000.

[4] Attaway. N. M., Bry. B. H."Parenting style and black adolescents' academic achievement", Journal of Black Psychology, vol 30,pp. 229–247, 2004.

[5] Aksenova S.S., Du Zhang & Meiliu Lu . "Enrollment Prediction through Data Mining". IEEE International Conference on Information Reuse and Integration,pp.510 – 515, Sept 2006.

[6] Barker, K., Trafalis, T., Rhoads, T. R."Learning from student data". Systems and Information Engineering Design Symposium,pp. 79-86, 2004.

[7] M. J. Druzdzel and C. Glymour. "Application of the TETRAD II program to the study of student retention in u.s. colleges ", In Working notes of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94), pp.419,1994.

[8] Herzog., S. "Estimating student retention and degree-completion time: Decision trees and neural networks Vis--Vis regression". New Directions for Institutional Research, pp.131,2006.

[9] Luna,J. "Predicting Student Retention and Academic Success at New Mexico Tech" , New Mexico Institute of Mining and Technology Socorro, New Mexico,2000.

[10] Massa, S. & Pulia_to, P.P. "An application of data mining to the problem of the university students' dropout using markov chains. In Principles of Data Mining and Knowledge Discovery", Third European Conference, PKDD'99, 51-60, Prague, Czech Republic,1999.

[11] Sanjeev, A.P. & Zytkow, J.M. "Discovering enrolment knowledge in university databases". In First International Conference on Knowledge Discovery and Data Mining, pp..246-251, Montreal, Que., Canada,1995.

[12] Song, Q., Chissom, B. S. "Forecasting enrolments with fuzzy time series", Fuzzy Sets and Systems, vol 62, pp.1-8,1994.

[13] Veitch, W. R. "Identifying characteristics of high school dropouts: Data mining with a decision tree model",2004.

[14] Bekele, R., & Menzel, W. "A bayesian approach to predict performance of a student (bapps): A case with ethiopian students". International Conference on AI and Applications,2005.

[15] Kotsiantis ,S. B., & Pintelas, P. E. "Predicting students marks in hellenic open university," IEEE Conference on Advanced Learning Technologies, vol. 30, pp. 664–668,2005.

[16] Martinez, D. "Predicting Student Outcomes Using Discriminant Function Analysis" .Annual Meeting of the Research and Planning Group. California, pp.163-173,2001.

[17] Superby, J.F. , Vandamme, J.P., Meskens, N. "Determination of Factors Influencing the Achievement of the First-year University Students using Data Mining Methods", Workshop on Educational Data Mining, pp.37-44,2006.

[18] Carneiro, P. “Equality of opportunity and educational achievement in Portugal”, Portuguese Economic Journal, vol 7 (1), pp. 17–41,2008.

[19] Yenilmez, K., Duman, A. ” Interviewing with students about the factors that affect the achievement of mathematic in Primary School”, Sosyal Bilimler Dergisi,vol 19, pp. 251–268,2008.

[20] Finn, J. D., Gerber, S. B., Achilles, C. M., Boyd-Zaharias, J. “The enduring effects of small classes. Teachers College Record,2000.

[21] Carpenter, P. “Single-sex schooling and girls’ academic achievements”, The Australian and New Zealand Journal of Sociology vol 21, pp.456-472, 1985.

[22] Arruabarrena,R., Pérez,T., López-Cuadrado, J., J Gutiérrez & Vadillo, J. (2002). “On Evaluating Adaptive Systems for Education”. Adaptive Hypermedia ,2347, 363-367 .

[23] Delavari, N., Beikzadeh, M. R., & Shirazi, M. R. A. “A new model for using data mining in higher educational system”. 5th International Conference. on Information Technology Based Higher Education and Training,2004.

[24] Shearer, C. “The CRISP-DM model: the new blueprint for data mining”. J Data Warehousing, pp13-22,2005.

[25] Kantardzic , M. “Data Mining: Concepts, Models, Methods, and Algorithms”, John Wiley & Sons,2003.

[26] Han, J. & Kamber, M. “ Data Mining:Concepts and Techniques, Second Edition“,University of Illinois at Urbana-Champaign,2001.

[27] Michalski, R. S., & Stepp, R. “Automated construction of classifications: Conceptual clustering versus numerical taxonomy”. IEEE Transaction on Pattern Analysis and Machine Intelligence (5), 396–410,1983.

[28] Berry, M. J. A., & Linoff, G. “Mastering data mining: The art and science of customer relationship management”. New York: John Wiley and Sons,2000.

[29] Quinlan J.R., “C4.5: Programs for Machine Learning”. San Mateo, CA: Morgan Kaufmann,1993.

[30] Romero, C., Ventura, S., Espejo, P. G., & Hervs, C. “Data mining algorithms to classify students”. 1st International Conference on Educational Data Mining, June 2008.

[31] Tjen S. Lim, Wei Y. Loh &Yu S. Shih. “A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms”. Machine Learning, Vol. 40, No. 3. pp. 203-228 ,2000.

[32] Bezdek, J. C. “Pattern Recognition with Fuzzy Objective Function Algorithms”. Plenum, New York,1981.

[33] Timofeev. R. “Classification and Regression Trees (CART) Theory and Applications”, a Master Thesis Presented ,2004.

[34] Atwell, R. H., Ding, W., Ehasz, M., Johnson, S., & Wang. M. “using data mining techniques to predict student development and retention”. In Proceedings of the National Symposium on Student Retention,2006.

[35] http://www2.cs.uregina.ca/~dbd/cs831/index.html

[36] Wu X, Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S., Zhou Z-H., Steinbach M., Hand D.J., & Steinberg D, “Top 10 algorithms in data mining”. Knowledge and Information Systems, vol 14, pp. 1-37,2008.

[37] Breiman L, Friedman J.H., Olshen R.A. & Stone C.J. “Classification and regression trees”. Wadsworth, Belmont,1984

[38] Delavari, N., Beikzadeh, M. R., & Shirazi, M. R. A. “A new model for using data mining in higher educational system”. 5th International Conference. on Information Technology Based Higher Education and Training,2004.

[39] Kotsiantis ,S. B., & Pintelas, P. E. “Predicting students marks in hellenic open university,” IEEE Conference on Advanced Learning Technologies, vol. 30, pp. 664–668,2005.