International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 2, No. 1, 2015, pp. 139-147.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Prediction of students' performance in high school by data mining classification techniques

Mohammadreza Zahedifard[a] , Iman Attarzadeh[b] , Hadi Pazhokhzadeh[c] , Javad Malekzadeh[d]

[a] *MS Student in Computer Engineering, Institute For Higher Education ACECR Khuzestan, Iran.*

[b] *Department of Computer Engineering, Islamic Azad University, Dezfoul Branch, Iran.*

[c] *MS Student in Computer Engineering, Islamic Azad University, Bushehr Branch, Iran.*

[d] *MS Student in Computer Engineering, University College of Rouzbahan, Iran.*

Abstract

Evaluation and prediction of students' performance in high school help to find important factors affecting students' success in education and moreover they can have an important role in helping educational managers in improving the quality of schools. According to this fact that data mining science has always been a suitable technique to extract knowledge from data, this can be used for providing a good approach. This article tries to present superior models in predicting students' performance. The mentioned data of this article are taken from 386 students of high schools in Bushehr province. In presented models best classification algorithms in educational data mining towards presenting superior predicting models have been provided. analyze and validation done on models shows the obtained results are precise and reliable. in this regard individual, environmental and educational factors affecting successful and unsuccessful students have been analyzed and according to them efficient models based on decision tree methods like c4.5 tree algorithm, support vector machine methods and logistic regression have been presented. The results can help managers of educational systems towards a correct educational planning an optimizing of educational processes in high schools.

Keywords: Educational data mining, Students' performance. Prediction, Classification techniques

Introduction:

Data mining or the method of knowledge extraction from data bases leads to discover new and useful patterns from large scales of information in data bases. specially in education field, these methods are used to investigate key questions and find empirical evidences for improving and developing of educational theories which in turn resulted in new domain of research called educational data mining. Educational data mining is used to addressing to different problems relating to education. according to the large scale of stored unprocessed data relating to students in education office and day to day increasing in this data, getting hidden knowledge from them is a necessity in improving educational quality and in a correct planning. hence , by using data mining techniques and based on individual, environmental and educational factors affecting on students performance we tried to make a useful model for them. a model in which ,individuals choose a correct planning based on their situations to improve the quality of their education and more than that managers of educational system by looking at the student's situations can have a better control and present educational advice to them and at the end by a correct educational planning help them to improve their educational quality. In this regard by collecting and analyzing the environmental and individual features of successful and unsuccessful people in second and third year of high school in different fields we extract the factors and properties affecting students' performance which has been neglected by educational managers, thereby we used data mining classification techniques to present suitable models in predicting successful or unsuccessful performance of high schools students. generally speaking, the purpose of this article is to present efficient models based on predicting techniques and by assuming parameters like individual and family features, student activity, personal features and... By using data mining classification algorithms.

In recent years, in order to optimize the large scales of stored data in educational systems and extraction of knowledge from this information, a lot of work has been done in this field. A comprehensive investigation of data mining researches done between 1995 to 2005 has been presented (romeo and ventura ,2007). checking and evaluation of different factors like sex features, the presence in classes, family income, and being a senior child and etc in leaving educational institutes has been analyzed by decision tree (quadri and kalyankar ,2010). students performance in learning environment along with affecting factors in improving it has been investigated (Kotsiantis et al ,2004). this study shows that student performance is not influenced only by their efforts but affected by other important factors(Bhardwaj and pal ,2012). data mining techniques has been used by crisp algorithm to design a recommender system based on educational performance records in order to support registration process of students (vialardi et al ,2011). by looking at students features and their educational records a Bayesian network model has been suggested (torabi et al ,2012). students information has been obtained by direct interview to find behavioral relations and effecting factors on their educational performance and then data mining techniques was used to find knowledge (oskouei and askari ,2014). incorrect use of educational environment by students and their irresponsible behavior was determined by classification techniques (baker and yacef ,2009). At the year of 2006, some studies on students' performance was done with this assumption that presence in class rate, hours of study after classes, income, family, age and level of mothers education are related to students performance (tahir and naqvi 2006). for improving learning predictors, bayes networks has been used (nandeshwar et al ,2011).

Data Collection

Since the objective of this paper is to predict students' performance in high schools based on individual, environmental and educational factors and according to this fact that stored information has a shortage in addressing individual and environmental factors , we decided to design a questionnaire with help of social and psychological experts and schools advisors in order to extract effecting features. The relationship of students' features and validation criterion has been shown in "Figure 1".

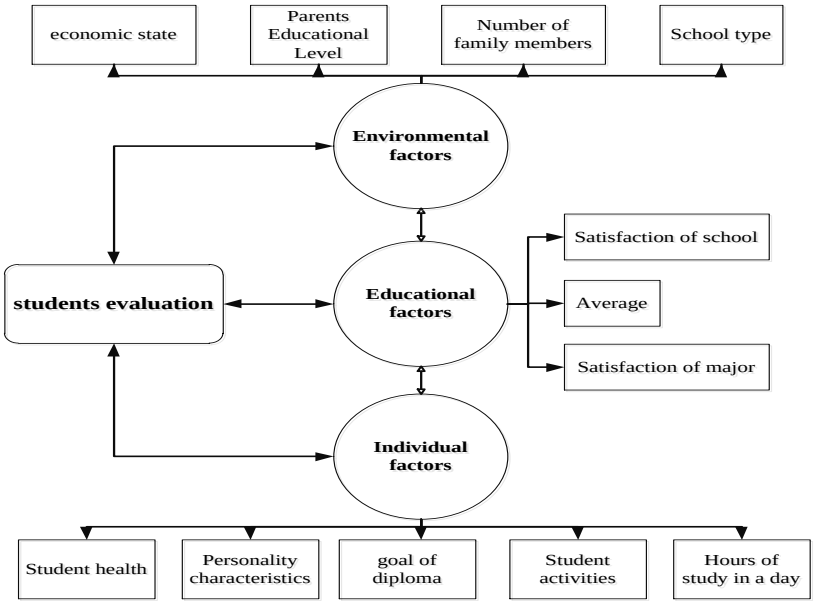


Fig (1) Collecting model of the effective factors on student

The collection of data used in this study are obtained from 386 successful and unsuccessful students of high schools which has been collected to predict students performance .information about students are collected separately in 5 field, mathematic and physic, natural science, humanities field, technical and occupational field and Work And Knowledge field, which every one of these files or categories has 69 predicting variables consisted of individual and family features, activities ,health and other personal features of student . Predicting models on each set of data are made according to the related variables. The number of variables in each data set is presented in "Table 1":

Table 1-information about variables and records of each data set

Number of records	Number of variables	Data set name
83	69	Mathamatic and physic
77	69	Natural science
62	69	Humanities
81	69	Technical and occupational
83	69	Work And Knowledge

Methodology

In order to apply data mining techniques to predict student performance in high schools by best data mining models the crisp algorithm was used. Crisp processing model has been designed in 1999 by 3 companies, Daimlerchrysler, SPSS and NCR (chapman et al 2000). this project method is a

standardization process of data mining in postindustrial use which are very interested among researchers. this method has got some exclusive features which one of them is its performance independent of any given domain. This means that this process can be used in any industry and application. Crisp methodology and process is more complete than SAMA methodology and according to the statistics and numbers it has better application, and more speed and possibility in success for execute in complicated business. this process includes 6 stages which these stages covers all required steps from first program to final evaluation .the suggested model presented in this study which has been taken from original crisp model are presented in "Figure 2":

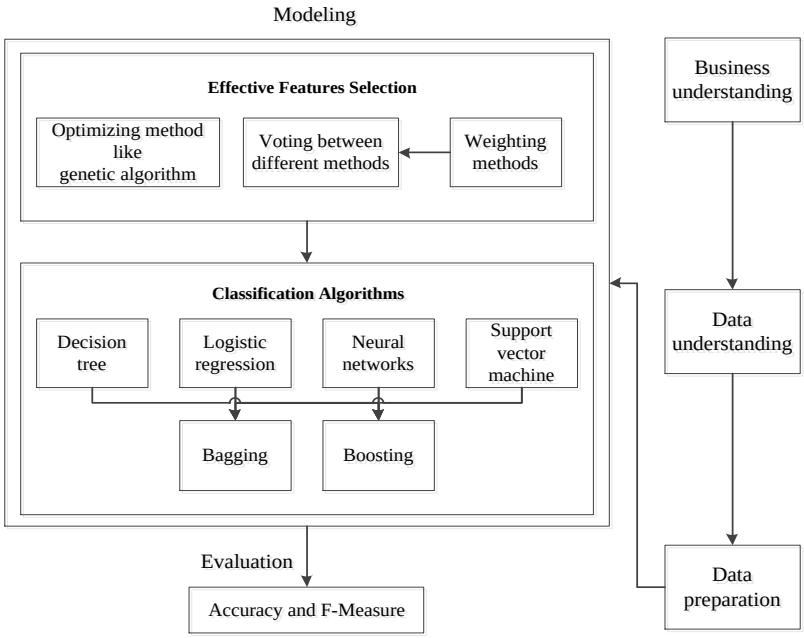


Fig (2) Proposed Model

Data preparation

At first, goal is to visualize data and make their important features and the relationship of variables clear. In diagram 1, box plot is used to show the relationship between two variables, shyness as a student features and GPA.

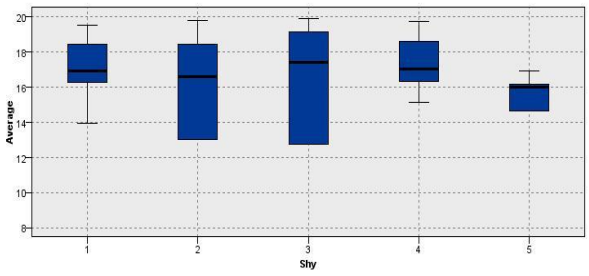


Fig (3) boxplot of 2 variables, average and student personal character, shyness.

According to the "Figure 3" above we can see that shy people with five rate point (very shy) have got a lower GPA comparing to other students. Students who are relatively shy have got highest GPA.
Before importing data to the modeling phase, data must be corrected. since data collection was done by questionnaire, number of lost quantities for variables was low, hence the correction of data hasn't been done on this stage. in primary selection stage, variables which no quantity hasn't been entered for them in more than 70% records or entered quantity was invariant for them in more than 70% of cases has been eliminated. The only target variable is student's average includes a few missing values. It's been decided to choose a value that provide the most frequency in the variable, and then replace new value for these missing values of variable. Students' success or failure based on GPA variable was determined by opinion of educational advisors of schools. Then target variable (GPA) for students with GPA higher than 17 was a successful situation and for students lower than 17 was a failure situation. All collected variables are discrete or became discrete, so as to, applying classification techniques can be done in a more precise and efficient. Since the number of variables is large, it was decided that some feature selection methods would be used to reduce variables. Since we had data set for 5 educational field ,selection of affecting features for every set was done. at this stage 3 group of feature selection has been used. first group feature selection methods are done based on weight, these groups includes GINI index, Information Gain, Chi Squared Statistic, Information gain ratio, Maximum Relevance and OneR. second method is done for feature selection among first group methods by voting and third method has been applied by optimizing method like genetic algorithm.

Results and evaluation

Difference in educational data and diversity in students requests in every educational system make educational researchers use different methods for finding hidden patterns in data bases. In this study, modeling is done by using 10-fold cross validation method which is one of the smartest methods among classification evaluation and along by different and best algorithms of classification. algorithms used at this stage are logistic regression methods, naïve Bayes, Bayes nets, nearest neighbor hood, decision tree (CART, C4.5, J48, CHAID and bayes), support vector machines, optimized support vector machines, artificial neural networks, bagging and boosting collective methods.in this regard all mentioned algorithms was applied to every each one of 5 data set. Then, only n variable with most weight has been chosen for modeling. in this study, n is a number between 8 to 40. after modeling, we should evaluate the performance of presented model in prediction of student s success or failure. The most important criterion in efficiency determination of an algorithm is accuracy index rating. this index shows what percent of all data sets was rated accurately. Calculation of accuracy index (taken from "Table 2") is shown in relation 1. (thammasiri et al,2014).

Table 2-Confusion matrix

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Accuracy index is not a useful index when dealing with data sets which have a unbalanced category with different number of samples. hence, F-Measure index which is a combination of precision and recall index is used that applies to cases in which we can't assume a special importance to precision and recall index. relation 2 shows the calculation method of F-Measure index.

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c} \quad (2)$$

In the following, best and superior models among all models (by using accuracy and F-Measure) on every data set
was shown.

First data set (mathematic and physic)
In "Table 3", superior models along with accuracy index and F-Measure are mentioned .as its clear, decision tree algorithm C4.5 along with feature selection method information gain ratio are presented as the most precise ones in this collection.

Table 3-results of modeling on first group of data

ID	FS Technique	Number of selected variables	Modeling method	Accuracy	F-Measure
Model1	Information gain ratio	10	C4.5	85.28	89.17
Model2	Information gain ratio	13	KStar	84.58	81.67
Model3	Information gain ratio	12	SVM	84.31	81.56
Model4	Information gain ratio	12	Bagging (SimpleCart)	84.31	82.26
Model5	Chi Squared	12	KStar	84.17	79.77

Second data set (natural science)
In "Table 4", superior models along with accuracy index and F-Measure for modeling on data set of natural science students are mentioned. as its clear logistic regression by using feature selection method chi-squared and selection of 10 effecting variables has been achieved a 92% accuracy in modeling.

Table 4-results of modeling on second group of data

ID	FS Technique	Number of selected variables	Modeling method	Accuracy	F-Measure
Model1	Chi Squared	10	multinomial Logistic Regression	92.14	94.49
Model2	Combination	8	linear Logistic Regression	92.14	94.49
Model3	Chi Squared	8	AdaBoost (SMO)	90.89	93.18
Model4	Combination	8	SVM	89.46	93.03
Model5	Chi Squared	16	Naive Bayes	88.21	91.15

Third data set (humanities)
In "Table 5", superior models along with accuracy index and F-Measure are presented for modeling on data set of humanities students. Like mathematics files data set, in this group decision tree methods by using a combination of feature selection methods are better than bayes network methods by a margin difference.

Table 5-results of modeling on third data set

ID	FS Technique	Number of selected variables	Modeling method	Accuracy	F-Measure
Model1	Chi Squared	11	Bayes Net	85.48	85.67
Model2	Combination	8	C4.5	84.52	84.02
Model3	Combination	8	CART	84.29	82.57
Model4	Chi Squared	11	SMO	83.81	83.88
Model5	Combination	8	Bagging (j48)	82.86	82.57

Fourth data set (technical and occupational field)
In "Table 6", superior models along with accuracy index and F-Measure for modeling on data set of technical and vocational students are mentioned. in these data group too, a decision tree established by 8 variables has been modeled by feature selection method information gain and achieved most precise accuracy by approximately 81% rate.

Table 6-results of modeling on fourth data set

ID	FS Technique	Number of selected variables	Modeling method	Accuracy	F-Measure
Model1	Information gain	8	Decision Tree	81.25	74.45
Model2	Chi Squared	16	SVM	0.8	73.34
Model3	Chi Squared	9	Neural Net	78.75	74.29
Model4	Chi Squared	8	NB Tree	78.75	75.52
Model5	Chi Squared	7	KStar	78.75	76.67

Fifth data set (work and knowledge field)
In "Table 7", superior models along with accuracy index and F-Measure for modeling on data set of these groups of student .in this data set too, like natural science data, logistic regression is the best method for modeling with a 88% accuracy.

Table 7-results of modeling on fifth data set

ID	FS Technique	Number of selected variables	Modeling method	Accuracy	F-Measure
Model1	Chi Squared	37	Logistic Regression	87.92	90.97
Model2	Chi Squared	23	Neural Net	85.14	88.86
Model3	Chi Squared	23	SVM	83.19	87.43
Model4	GA	-	Neural Net	80.83	85.57
Model5	Combination	10	Bagging (KStar)	80.69	85.06

In the following, "Table 8" relating matters to prediction in educational data mining along with obtained accuracy has been shown.

Table 8-accuracy of related models in the field of educational data mining

Authors	Year	Number of records	Data resources	Accuracy
E Osmanbegović et al	2012	257	University of Tuzla	76.65
LE Livieris et al	2012	279	Lyceum in Greece	87.1
Y Zhang et al	2010	4223	University	89.5
DorinaKabakchieva	2013	10330	Bulgarian university	85.00
DursunDelen	2010	16066	Various University	87.23
ZJ Kovačić	2010	453	University	60.5
Thammasiri et al	2014	21654	university	90.24
S Sembiring et al	2011	1000	UMP	93.67

According to the table 8 and by evaluation and comparison of same models we conclude that all models with 90% accuracy have got more than 1000 samples. so, due to the limited collected records for every data group, obtained accuracy of presented models are remarkable. Generally speaking, evaluation of 5 model of every data set in this study shows that weight-assignment methods to variables for choosing feature selection is a better method than feature selection based on optimizing like genetic algorithm. Also among weight-assignment methods, chi-squared method almost in every data set can choose most effective variables. As a general assessment of superior classification models we can say that methods based on decision tree like c4.5 tree, support vector machine and logistic regression were best methods in classification. Also it is clear that in all data set, by choosing at least 8 to 23 affecting variables among all variables most precise in modeling can be achieved. A precise estimation of future proceeding of students for every educational system is a necessity. obtained results can improve quality of education and learning. Advisors and teachers can take an important contribution in these results and encourage students to participate in different activities to improve their performance. In addition, managers also can use the results of this classification to prioritize educational procedures and improve resource allocation. if we collect more data from schools data base, a more precise results can be achieved .because a low data volume can affect the resulted accuracy.

References:

Anupama Kumar, S. Vijayalakshmi, M. N. (2012). Mining of student academic evaluation records in higher education. In IEEE International Conference on Recent Advances in Computing and Software Systems (RACSS). Chennai. India. 25 Apr - 27 Apr. 67-70

Baker, R. S. Yacef, K.)2009(. The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining. Vol.1. No.1. 3-17

Bhardwaj, B. K. Pal, S.)2012(. Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security (IJCSIS). Vol. 9. No. 4. 136-140

Chapman, P. Clinton, J. Kerber, R. Khabaza, T. Reinartz, T. Shearer, C. Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS

Delen, D. (2010). "A comparative analysis of machine learning techniques for student retention management", Decision Support Systems, Vol.49. No.4. 498-506

Expert Systems with Applications 38(12). 14984-14996

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies. Vol.13. No.1. 61-72

Kotsiantis, S.B. Pierrakeas, C.J. and Pintelas, P.E. (2004). Predicting Students Performance in Distance learning Using Machine Learning Techniques. Applied Artificial Intelligence. Vol.18. No.5. 411-426

Livieris, I. E. Drakopoulou, K. Pintelas, P. (2012). Predicting students' performance using artificial neural networks. In Proceedings of the 8th Pan-Hellenic Conference Information and Communication Technology in Education. Jelgava. Latvia. April 23-24. 190-193

Nandeshwar, A. Menzies, T. and Nelson, A. (2011). Learning patterns of university student retention.

Oskouei, R. J. Askari, M.) 2014(. Predicting Academic Performance with Applying Data Mining Techniques (Generalizing the results of two Different Case Studies). Computer Engineering and Applications Journal. Vol.3. No.2. 79-88

Quadri, M. Kalyankar, D. N.)2010(. Drop out feature of student data for academic performance using decision tree techniques. Global Journal of Computer Science and Technology. Vol.10. No.2. 2-5

Romero, C. Ventura, S.)2007(. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications. Vol.33. No.1. 135-146

Sembiring, S. Zarlis, M. Hartama, D. Ramliana, S. Wani, E. (2011), Prediction of student academic performance by an application of data mining techniqes. International Proceedings of Economics Development & Research, Vol.6. 110-114

Tahir, S. Naqvi, S. R. (2006). Factors Affecting Students' Performance A Case Of Private Colleges. Bangladesh e-journal of sociology. Vol.3. No.1. 90-100

Thammasiri, D. Delen, D. Meesad, P. Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications, Vol.41, No.2, 321-330

Torabi, R. Moradi, P. Khantaimoori, A. R.)2012(. Predict student scores using bayesian networks. Procedia-Social and Behavioral Sciences. Vol.46. 4476-4480

Vialardi, C. Chue, J. Peche, J. P. Alvarado, G. Vinatea, B. Estrella, J. Ortigosa, Á.)2011(. A data mining approach to guide students through the enrollment process based on academic performance. User modeling and user-adapted interaction. Vol. 21. No.1-2. 217-248

Yadav, S. K. Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal (WCSIT). Vol. 2. No. 2. 51-56

Zhang, Y. Oussena, S. Clark, T. Hyensook, K. (2010). Using data mining to improve student retention in HE: a case study. 12th International Conerence on Enterprise Information Systems(ICEIS). Madeira. Portugal. June 8-12. 1-8