RESEARCH PAPER | Mathematics | Volume : 5 | Issue : 3 | March 2015 | ISSN - 2249-555X

Assessment of math score using statistical techniques

KEYWORDS: ……..; logistic regression analysis; neural network; discriminant analysis.

Ilyas Akhisar
Marmara University, School of Banking and Insurance, Goztepe Campus, Istanbul TURKEY

Abdulkadir Tepecik
Yalova University, Engineering Faculty, Sehit Omer Faydali Cad. Yalova 77100 TURKEY

ABSTRACT *Understanding the factors that refer math scores students an interesting and curious problem. Acknowledge is hidden among the educational data set that is extractable using data mining techniques. In this research, Statistical techniques are used to evaluate students' performance and as there are many approaches that are used for data classification is used. We address this issue by investigating of two groups statistically testing. In the process, two different math teaching methods applied to both groups. In this paper, which factors are highly effect to students' math score are investigated and results interpreted.*

INTRODUCTION

Recently, rapid development of information technologies increased the amount of the collected data and challenges associated with managing and analyzing data.

Data Mining also popularly known as knowledge invention in database is to provide information and to become aware of knowledge from huge data set (Baradwaj and Saurabh, 2011).

DM seeks the ways that organize information about the data structure of hidden relationships and partnership rules, the unknown elements representing the estimated useful results to classify objects and to decide (Ayala, 2014).

Recently, the DM design models reveal for training, tasks, methods and algorithms based on data (Anjewierden et al, 2007).

Educational Data Mining (EDM), the view out to find a way to characterize and predict the success of students and evaluation of educational applications (Anjewierden et al, 2007).

Understanding the factors that conduct to success (or failure) of students at public secondary education is a difficult issue. Therefore, determining the variables related to success of students have always been arose the curiosity of researches (Sen et al 2012).

The aim of this study is that using different DM methods to determine the predictive variables (i.e. factors) by applying to data. Moreover discovering variables that are related with the achievements would be beneficial to students, parents, teachers, administers. In addition the results of the study would be valuable for researchers and practitioners.

This study is organized as follows literature survey, Turkish education system, research methodologies and interpretation of results. In the last section, concluding remarks are given.

In literature, the Educational Data Mining (EDM) community website, www.educationaldatamining.org concerned with developing methods for exploring data that come from the educational setting, and utilize some methods to better understand students behavioral pattern of learning (Siti and Tasir, 2013).

Regression technique is the act of finding a relationship between one or more independent variables and dependent variable on it.

Independent variables and dependent variables that are to be predicted in data mining is already known values. However many world problems are not simply guess. Therefore to estimate the actual values it may be necessary more complex techniques such as Logistic Regression (LR).

Hijazi and Naqvi handled a study of the yield of selecting 75 females of 300 students from colleagues of Punjab University in Pakistan.

The hypothesis that was built as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, parents' age and education are significantly related with student performance". By simple linear regression analysis shown that parents' education and income are highly correlated with the student academic performance (Hijazi and Naqvi, 2006).

Levy and Wilensky used LR to investigate students'survey responses of relating physical variables of the system (Levy and Wilensky, 2011).

Classification is the most frequently to put in practice in data mining technique works for developing a model that can classify the population of records. This approach frequently employs neural network-based classification is a set of connected input/output units and each weighted connection. Throughout the learning phase, network learns by adjusting weights to be able to estimate the correct class labels of the input tuples.

Driving a remarkable sense from obscure data and acquiring patterns and perceiving trends which are not having knowledge by computer techniques as well.

Wang and Liao to find out the recent learning performance of students to predict future performance in learning English using back propagation in Neural Networks (NN) to predict the classification (Wang and Liao, 2011).

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for

Volume : 5 | Issue : 3 | March 2015 | ISSN - 2249-555X

proper discrimination. The classifier is encoding algorithm parameters into a model. If accuracy is admissible, the rules can be applied to the new data.

Kock and Paramythis using Linear Discriminant Analysis (LDA) to find clustering to represent learners' to detect problem solving styles (Kock and Paramythis, 2011).

The cluster is selected for further analysis is a sampling selection portioned into groups and random sample of the population.

The research shown that the girls with high and the boys with low socio-economic status relatively correlated of the academic achievements in the area of science stream.

Khan, conducted a performance study on main objective to demonstrate the prognostic value of different measures of cognition, personality and demographic variables for success at higher in science stream for 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University (Pandey and Pal, 2011).

R.Martinez, et al. used hierarchical agglomerative clustering to extract patterns of activity for unveiling the strategies followed by groups of learners (Martinez et al, 2011).

Ayesha et al, depicts the use of k-means clustering algorithm is usually used to objective classification to see students' learning activities (Shaeela et al, 2010).

Clustering is to identify of similar classes of objects, by using clustering techniques to discover over all pattern and relations of data (Baradwaj and Saurabh, 2011).

MATERIALS AND METHODS

Students learn math depending on maturity, mental abilities, experience, performance, preferred learning styles, attitudes mathematics.

Teachers to teach trials depends math, according to their understanding beliefs and mathematics itself, their own preferred styles with students and faculty, children's views about the role assessment, professionalism and effectiveness of mathematics as a teacher.

Data experimental class (30 students) and control class (30 students) in the secondary school 6th class with different sections of public revenues.

Table 1 .Data

Variables	Divisions
Mother Education Level (x_1) Father Education Level (x_2)	Primary School Secondary School High School Under Graduate Graduate
Mother Job (x_3) Father Job (x_4)	Civil Servant Worker Self-Employment Other
Mother Revenue (x_5) Father Revenue (x_6)	Less than 500 TL (TL: Turkish Lira) > 500 TL and <1000 TL >= 1000 TL and < 2000 TL >= 2000 TL and < 4000 TL >= 4000 TL
Who Helps to Student (x_7)	Mother Father Brother(s) and/or Sister(s) Other(s) None
When Helps to Student (x_8)	Never Rarely Sometimes Frequently Always
Student Work (x_9)	Yes No
Tutoring (x_{10})	Yes No

One of the analytical methods is LR model that is the relationship between a set of independent variables and the probability that a case is a member of one of the categories of the dependent variable.

The second method in NN architectural structures, nodes are organized into groups called layers. Input layers receive inputs, output layers produce outputs and internal (or hidden) layers provide the interconnections between input and output.

The NN provide a tool for describing non-linearity in volatility processes of financial data and help to answer the question of "how much" non-linearity is present in the data (Miazhynskaiaa et al, 2006).

Last method, LDA is a well-known classical statistical technique to find the projection that maximizes the ratio of scatter among the data of different classes to scatter within the data of the same class. In the late 1960s, LDA was introduced to create an empirical indicator of financial ratios. Beaver using financial ratios developed on indicator that best differentiated between failed and non-failed firms using univariate analysis techniques (Beaver, 1966).

The univariate approach was later improved and extended to multivariate analysis by Altman, considered several variables simultaneously using multiple discriminant analysis (MDA). During the next years that followed, many researchers attempted to increase the success of MDA in predicting business failure (Dimitras et al 1996).

RESULTS

In this study, the application of LR model to data, we can have classification table and variables in equation when we inquire the Table 1. The classification table, control group and the experimental group which is applied new teaching method and evaluating/grading (i.e. Group1 and Group 2) over all predicted correct percentage moreover than sixty.

Volume : 5 | Issue : 3 | March 2015 | ISSN - 2249-555X

Table 2. Classification table

Observed		Predicted			
			Group_1_2		Percentage Correct
			1	2	
Step 3	Group_1_2	1.00	22	8	73.3
		2.00	11	19	63.3
	Overall Percentage				68.3
The cut value is , 500					

The LR model as follow

$$\ln\left(\frac{p}{1-p}\right) = -0.921 + 0.649 * x_2 + 0.538 * x_8$$

where one unit increase the level x_2 up contributes 1.914 times and one unit increase the level of x_8 up contributes 1.713 times the students' Math

Table 3. Score variables in equation

	B	Wald	df	Sig.	Exp(B)
x_2	.649	3.590	1	.058	1.914
x_8	.538	4.888	1	.027	1.713
Constant	-.920	.802	1	.371	.398

Table 4. Step Summary

Step	Model			Correct Class %	Variables
	Chi Square	df	Sig		
1	6.613	1	.01	61.7 %	IN: x_8
2	15.510	3	.01	68.3 %	IN: x_2
a. No more variables can be deleted from or added to the correct model					

In LR results, the step summary table shows that both the variable, x_8 and x_2 are statistically significant, others are not.

Statistical technique that can classify the population records by using NN based algorithm.

The NN results show that especially x_2 maximum importance to students Math Score than others. On the other hand x_8 variable more influence many other variables as well as LR results.

Table 5. Independent Variable Importance

	Importance	Normalized Importance
x_1	.079	58. 2%
x_2	.136	100.0%
x_4	.110	80. 7%
x_3	.107	78. 6%
x_6	.063	46. 2%
x_5	.131	76. 8%
x_9	.078	57. 1%
x_{10}	.127	93. 3%
x_7	.070	51. 3%
x_8	.101	84. 2%

Table 6. Model Summary

Training	Sum of Squares Error	14.527
	Percent Incorrect Predictions	31.8%
Testing	Sum of Squares Error	16.946
	Percent Incorrect Predictions	34.8%

The neural network model training and test testing to predict the correct percentages are almost adequate to represent the model.

Application of LDA technique, the eigenvalues table show that canonical correlation of model is 0, 981 and a linear discriminant function discriminate function discriminate groups.

Table 7. Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Corr.
1	25.88	100.0	100.0	.981

Eigenvalues table is analysed that the model explain (0, 981)2=0, 96 % of the total knowledge.

Table 8. Canonical Discriminant Function Coefficient

	Function
	1
x_2	2.375
x_6	-1.705
(Constant)	-2.290

$$z = -2.290 + 2.375 * x_2 - 1.705 * x_6$$

The categorical dependent variables (two groups) discriminate by using linear discriminant function of the model is give above. According to model that is a positive correlation between students 'Math Score and the students whom x_2. The student Math Score is 2, 375 times more than when a students' x_2 is up one class according to the other student's x_2 just below class. On the other hand there is a negative correlation the Math Score and The Students' x_6.

CONCLUSION

In this paper, the classification task is used on student database to predict the students' math score. As there are many approaches that are used for data classification.

As the results indicate, all of the classification methods performed reasonably well in predicting reasonable factors which the father education level of the student and how often helped to the class of student factors arose in statistical techniques which are considered. Family background and social-economic status are critical for student's math score is obtained as the results of this research.

This study will help the students, teachers, administers and parents. In addition the results of the study would be valuable for researchers and practitioners.

Acknowledgements:

Volume : 5 | Issue : 3 | March 2015 | ISSN - 2249-555X

REFERENCE

1. Ayala A.P. (2014) Educational data mining: A survey and a data mining-based analysis of recent works Expert Systems with Applications, 41, 1432–62. | 2. Anjewierden A, Kolloffel B , and Hulshof, C. (2007) Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes, In Proceedings of the international workshop on applying data mining in e-Learning, 23–32. | 3. Luan J, (2002) Data mining and its applications in higher education, Journal of New Directions for Institutional Research, 113, 17–36. | 4. Sen B, Ucar E and Delen D (2012) Predicting and analyzing secondary education placement-test scores: A data mining approach, Expert Systems with Applications 39, 9468–76. | 5. Siti K. M, Tasir Z (2013) Educational data mining: A review, Procedia - Social and Behavioral Sciences 97 320–24. | 6. Hijazi S T, Naqvi R S M M (2006) Factors affecting student's performance: A Case of Private Colleges, Bangladesh e-Journal of Sociology, 3(1). | 7. Levy S T , Wilensky U (2011) Mining students' inquiry actions for understanding of complex systems Computers & Education, Vol.56(3), 556–73. | 8. Wang Y H , Liao H C (2011) Data mining for adaptive learning in a TESL-based e-learning system Expert Systems with Applications, 38(6), 6480–85. | 9. Kock M , Paramythis A (2011) Activity sequence modelling and dynamic clustering for personalized e-learning, User Modeling and User-Adapted Interaction, 21(1–2) 51–97. | 10. Pandey U K, Pal S (2011) A Data mining view on class room teaching language", International Journal of Computer Science Issue(IJCSI), 8(2), 277-82. | 11. Martinez R, Yacef K, Kay J , Al-Qaraghuli , Kharrufa A (2011) Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop In Proceedings of the 4th international conference on educational data mining 111–20. | 12. Shaeela A, Tasleem M, Sattar A R , Khan I (2010) Data mining model for higher education system, European Journal of Scientific Research, 43(1). | 13. Baradwaj B K, Saurabh P (2011) Mining Educational Data to Analyze Students" Performance, International Journal of Advanced Computer Science and Applications (IJACSA), 2(6). | 14. Miazhynskaiaa T, Frühwirth-Schnatter S, Dorffner G (2006) Bayesian testing for non-linearity in volatility modeling, Computational Statistics & Data Analysis | 51 2029–42. | 15. Beaver W (1966) Financial Ratios as Predictors of Failure, The Journal of Ac- | counting Research 4 71-102 | 16. Dimitras A I, Zanakis S H, Zopounidis C (1996) A Survey of Business Failure with an Emphasis on Prediction Methods and Industrial Applications, European Journal of Operational Research 90 487-513 | 17. Hyun C K, Daijin K, Sung Y B (2003) Extensions of LDA by PCA Mixture Model and class-Wise Features, The Journal of Pattern Recognition 36 1095-1105 | 18. Fukunaga K, Short R D (1980) A Class of Future Extraction Criteria and Its relation to the Bayes Risk Estimate, IEE Transaction an Information Theory, IT-26