A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh

Mohammad Nayeem Hasan
Department of Statistics Shahjalal
University of Science and Technology
Sylhet, Bangladesh
nayeem5847@gmail.com

***Abstract*—This study aimed to compare the predictive accuracy and also the classification accuracy of two models using real data of school attrition. The overall classification accuracy for both models was determined by the classification accuracy rate. Logistic regression analysis (LRA) and linear discriminant analysis (LDA) classified 78.33% and 78.38% of girls respectively, in-school and out-of-school correctly. The AUROC curve for LRA was 80.63%, while it was 80.57% for the LDA. The LRA has sensitivity and specificity were 45.81% and 91.60%, respectively, and the LDA had a sensitivity of 46.81% and specificity of 91.01%. The overall classification rate for both was good. In comparison with the conventional LRA model, the LDA was better than LRA in the correct classification rate. In general, the LRA model looks appropriate for prediction accuracy while LDA seems suitable to be used for classification techniques.**

Keywords*—school dropout, classification accuracy, LDA, LRA

I. Introduction

In the present world, education is called investment [1]. It is the dynamic energy behind any emerging economy. It creates the possibility and delivers to society by a well-educated and skilled man forces [2]. Girls are vital for the progress of any country by teaching the fundamental use of education, well-educated women have better income and fewer hesitation, healthier and enhanced educated children compare with more than the boys [3]. But for decades women in Bangladesh have been discriminated against in education.

Nowadays, the proportion of girls' attrition from school remains high in many low-income nations, in the figure, more than half of the girls start primary education but couldn't finish it [4]. As compared to boys the attrition rate is higher for girls in 49 countries [5]. A student's family poverty level effect on student attrition behavior [6]. School-to-school shifts are one of the reasons behind the secondary school attrition rate [7]. School's academic and social climate influenced institutional attrition ratios [6].

For classifying/predicting the binary outcome variable (attrition), some methods are available such as linear discriminant analysis, logistic regression analysis, and different data mining models were also available. Several studies have been conducted to identify the predictors of school attrition by constructing a bivariate and multivariate logistic regression model. Several studies have been conducted to compare different classified models of different fields. From that, many studies specified that LDA is more powerful for classifying a variable into several groups, where another confirmed LRA is suitable in predicting and others concluded that both models had a similar performance [8]-[9].

The purpose of this work, after summarizing the characteristics of the two discriminatory methods, is to find out the transformation of the two analytical approaches when pediatric educational researches are used to assess attrition outcomes. In particular, we applied a logistic regression model to predict the attrition risk based on effective factors and then we have compared the predictive accuracy and also the classification accuracy of each model using real data of school attrition of female students in Bangladesh.

II. Data and Methodology

A. Study Design

Data were used from the latest available the Multiple Indicator Cluster Survey (MICS 2012) data. The cross-sectional survey is one of the largest surveys conducted in Bangladesh, which is based on a sample of 51895 households (43474 rural, 8421 urban) interviewed with a response rate of 98.5%.

B. Response Variable

In this study, the response variable is the school attrition. Attrition is identified by a woman, was ever attended school but not in school in the current school year before the data collection Table I.

C. Predictor Variables

Several factors are associated with the girl's school attrition. Predictor variables based on the previous study are [5], [8], [9]– [10] included in this study Table I.

TABLE I. Factors Used for Predicting School Attrition

Predictor Variables	Response Variable
Girls age (Year)	School attrition
Marital status	
Area	
Divisions	
Household wealth index	
Household education	
Religion	
Mother alive	
Father alive	

978-1-7281-6040-5/19/$31.00 ©2019 IEEE

D. Data Mining Approach

The predictive accuracy for both models was estimated by the area under the receiver operating characteristics (AUROC) curves Table II. Higher AUROC indicated a better-predicted accuracy of the models. The total classification rate for both models was determined by comparing the predicted events with the actual events Table III. The data management and data analysis for this study has been carried out using R.

III. Results

A total of 4800 women generated as a subsample. Among them 15 (36.40%), 16 (34.10%) and 17-year-old (29.50%) women were included in the analysis, respectively (Fig. 1).

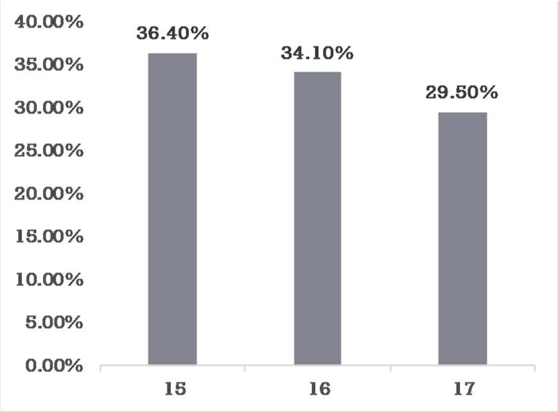

Fig. 1. Percentage of girls according to age category

Among 15-year-old women, 7.20% were out of school and 29.20% of women were attending school. Similarly, the percentage of 16 and 17-year-old women were 10.60% and 11.30%, respectively (Fig. 2).The template is designed for, but not limited to, six authors.

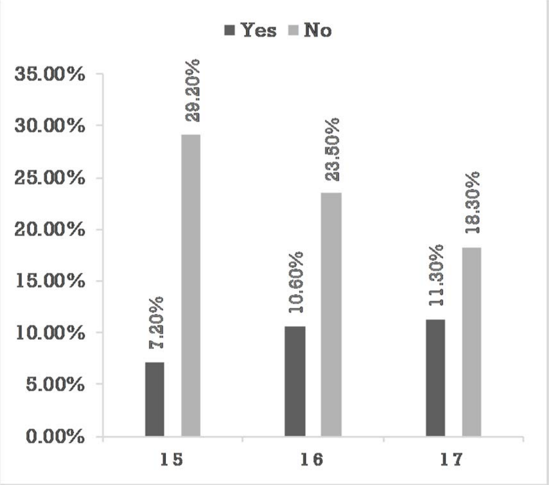

Fig. 2. Percentage of girls by age and attrition

In Fig. 3, the Sylhet division had the highest attrition among girls age between 15-17. The results indicated that about 44.47% of girls for these regions were a dropout. The lowest percentage detected in Barisal, 22.18 % and for Chattagram, Dhaka, Khulna, Rajshahi and Rangpur charges of attrition remained 30.87%, 28.83%, 24.93%, 29%, and 2.42% respectively.

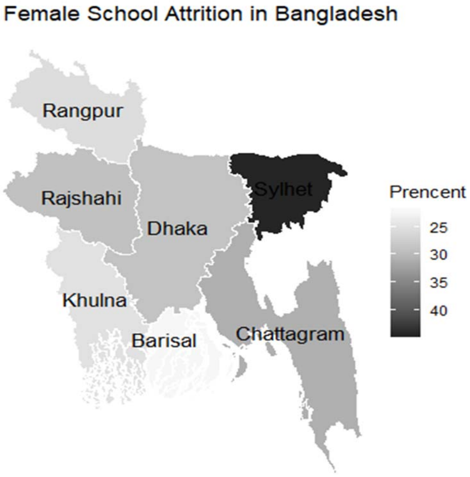

Fig. 3. Division wise School Attrition in Bangladesh

In Fig. 4, variable importance measure marital status, wealth index, division, the age of women and household education were the first five effective factors on school attrition, respectively.

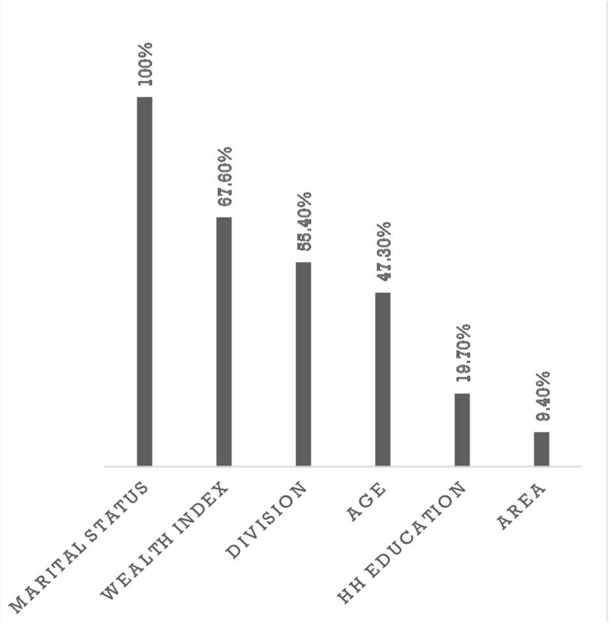

Fig. 4. Factors affecting school attrition in the order of their importance from school

The AUROC for LRA and LDA models separately were compared with the reference AUROC. The AUROC curve for LRA was 80.63%, while it was 80.57% for the LDA. The sensitivity and specificity of LRA were 45.81% and 91.60%, respectively, and the LDA had a sensitivity of 46.81% and specificity 91.01%, respectively Table II.

TABLE II. Test of Sensitivity, Specifity, and AUROC for LRA and LDA

Model	Sensitivity (%)	1-Specificity (%)	AUROC
LRA	45.81%	91.60%	80.63%
LDA	46.81%	91.01%	80.57%

Also, the kappa statistics for LRA were 0.4159 and this statistic for the LDA was 0.4230. The LRA and LDA models classified 78.33% and 78.38% of students respectively, in-school and out-of-school correctly Table III.

TABLE III. TEST OF ACCURACY AND KAPPA STATISTICS FOR LRA AND LDA

Model		Yes	No	Accuracy (%)	Kappa Statistics
LRA	**Yes**	638	283	78.33%	0.4159
	No	757	3122		
LDA	**Yes**	661	304	78.38%	04230
	No	734	3101		

IV. DISCUSSION

In this study, the LDA model and LRA model were compared using the overall classification accuracy and comparing the prediction accuracy using AUROC.

We conclude that this information has given similar results in both LRA and LDA performance. Overall classification rates were good for both and could be helpful in either classification of the class of attrition. In our study, LRA slightly exceeds LDA in the correct classification rate and LRA performed better than LDA. In a paper, K. Kitbumrungrat examined the accuracy of LDA and LRA in breast cancer data and get a similar result of this study [10]. But when taking into account sensitivity, specificity and AUROC, the differences in the AUROC were negligible. It has been shown that many types of data are found similar results [5], [11].

G. Kwame Abledu exposed that LDA was more precise than LRA for small trials [12]. M. Pohar, M. Blas, and S. Turk presented that sample size had an acute effect on making a decision on the difference between models [13]. Therefore, the size of the sample had some effect on the accuracy of the classification, although a smaller sample size had more effect on the LRA than the LDA.

Discriminant analysis can be used if there are usually more than three or more groups as a categorical variable in the dependent variables. The number of predictions in the discriminant analysis is equal to the number of categories in the variable less than one of those categories. Thus, for n categories, the response variable presents the n-1 equations, and one or two have the required strength to attain the best classification accuracy. In this case, the complexity becomes about the number of equations that need to be taken from the existing set of equations [8]. Thus, the researchers conclude that the choice of LDA is higher for initial or stepwise analysis, otherwise, LRA should be used.

V. CONCLUSION

The objective of this study, to appropriate the use of LDA and LRA models to examine the classification/prediction of school attrition in Bangladesh. The models were used to classify the attrition variable as in-school and out-of-school. From the predicting performance and classification accuracy of LDA and LRA, we decided that the overall classified rates of these two models were acceptable and it could be acceptable in classifying/predicting the school attrition.

Finally, the choice of the appropriate model selection method depends on where the model needs to be applied, along with the probability of the necessary assumptions. In conclusion, deciding which method to use, we must consider the assumptions applied to each one.

VI. ETHICS APPROVAL

This study involves secondary data analyses of a publicly available dataset that are freely available on the UNICEF website with all identifier information removed, ethical approval from respective institution was not required.

VII. CONCLUSION

The objective of this study, to appropriate the use of LDA and LRA models to examine the classification/prediction of school attrition in Bangladesh. The models were used to classify the attrition variable as in-school and out-of-school. From the predicting performance and classification accuracy of LDA and LRA, we decided that the overall classified rates of these two models were acceptable and it could be acceptable in classifying/predicting the school attrition.

Finally, the choice of the appropriate model selection method depends on where the model needs to be applied, along with the probability of the necessary assumptions. In conclusion, deciding which method to use, we must consider the assumptions applied to each one.

REFERENCES

[1] Okumu et.al, "Socioeconomic Determinants of Primary School Dropout: The Logistic Model Analysis1 By," no. 7851, 2008.

[2] G. C. Gondwe, "Factors Influencing Rural Female Pupils Drop Out from Primary Schools , in Nkhata-Bay South District , Malawi," 2016.

[3] M. A. Amadi, E. Role, and L. N. Makewa, "Girl Child Dropout: Experiential Teacher and Student Perceptions," Int. J. Humanit. Soc. Sci., vol. 3, no. 5, p. 8, 2013.

[4] R. Sabates, A. Hossain, and K. M. Lewin, Consortium for Research on Educational Access, Transitions and Equity School Drop Out in Bangladesh: New Insights from Longitudinal Evidence, no. 49. 2010.

[5] A. H. M. Shahidul, S. M. and Karim, "Factors Contributing to School Dropout Among the Girls," Eur. J. Res. Reflect. Educ. Sci., vol. 3, no. 2, pp. 25–36, 2015.

[6] P. Goldschmidt and J. Wang, "When can schools affect dropout behavior? A longitudinal multilevel analysis," Am. Educ. Res. J., vol. 36, no. 4, pp. 715–738, 1999.

[7] J. W. Alspaugh, "The Effect of Transition Grade to High School, Gender, and Grade Level Upon Dropout Rates," Am. Second. Educ., vol. 29, no. 1, pp. 2–9, 2000.

[8] Z. Shayan, N. Mohammad Gholi Mezerji, L. Shayan, and P. Naseri, "Prediction of Depression in Cancer Patients With Different Classification Criteria, Linear Discriminant Analysis versus Logistic Regression.," Glob. J. Health Sci., vol. 8, no. 7, pp. 41–6, Nov. 2015.

[9] C.Y. Liong and S.F. Foo, "Comparison of linear discriminant analysis and logistic regression for data classification," 2013, pp. 1159–1165.

[10] K. Kitbumrungrat, "Comparison Logistic Regression and Discriminant Analysis in classification groups for Breast Cancer." 2012.

[11] M. Nyoni, T. Nyoni, and B. Wellington Garikai, "Factors Affecting Students' Academic Achievement in Zimbabwe's Rural Secondary Schools: A Case Study of Marimasimbe Secondary School in Jiri Community," Dyn. Res. Journals' J. Econ. Financ., vol. 2, pp. 1–15, 2017.

[12] G. Kwame Abledu, "Comparison of Logistic Regression and Linear Discriminant Analyses of the Determinants of Financial Sustainability of Rural Banks in Ghana," Am. J. Theor. Appl. Stat., vol. 5, no. 2, p. 49, Mar. 2016.

[13] M. Pohar, M. Blas, and S. Turk, "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study," 2004.