Eur. J. Stat. 3 (2023) 8
doi: 10.28924/ada/stat.3.8

# Application of Discriminant Analysis to Predict Students' Performances in Mathematics in Advanced Secondary Schools

Nuhu Saidi, Gadde Srinivasa Rao*

*Department of Mathematics and Statistics, College of Natural and Mathematical Sciences, University of Dodoma, P.O. Box 259, Dodoma, Tanzania*
*milawandu123@gmail.com, gaddesrao@gmail.com*
**Correspondence: gaddesrao@gmail.com*

Abstract. This quantitative study aimed to use discriminant analysis procedures, to develop a classification model to be used for prediction, to predict students' performances in Mathematics in advanced secondary schools in Tanzania. The study was conducted in Iringa Rural District to model students' performances in Mathematics in advanced secondary schools owned by the government. Secondary data of students' performances in Mathematics of 126 students when they were form five in the year 2020/2021 were collected from academic students' progressive reports and three distinct groups each contained 42 students' performances were formed. The analysis was done by using R programming software and a seed of 66 was used during the data partitioning to create training and test datasets. The maximum posterior probability rule was used as a classification rule to assign students' performances in Mathematics into three proposed groups which are: High, Medium and Low. The classification accuracy achieved by the classification model to classify students' performances in the training dataset is 97.33%. During validation, the model achieved the classification accuracy of 96.08% to classify students' performances in the test dataset. These findings imply that, the classification model is valid and reliable. Hence the model is convenient to be used for prediction, to predict students' performances in Mathematics in Advanced Certificate of Secondary Education Examinations.

## 1. Introduction

The problem of poor performance in Mathematics is very common all over the world, but the problem is more severe in developing countries [15]. Tanzania is one among the developing countries that has been experiencing the problem of poor performance in Mathematics both in primary and secondary schools [13]. In Tanzania, many students are trapped in a vicious circle of poor performance in Mathematics and the problem does not confine only to Mathematics, but it also extends further to other subjects, particularly science subjects which also need good foundations of Mathematics. The problem is more severe due to various reasons such as poor infrastructures, lack of teaching and learning materials, few numbers of Mathematics teachers and

Received: 11 Dec 2022.
*Key words and phrases.* classification model; discriminant function; classification rule; classification accuracy.

others [18]. So, despite the importance of mathematics in many disciplines, still many students find it difficult to pass and fail to pursue some courses in tertiary education which need good foundations of Mathematics.

Due to the significant contribution of Mathematics for the bright future of the country's development, particularly in the fields of engineering and technology, many researchers in Tanzania conducted numerous studies to find the ideal solutions to overcome this problem of poor performance in Mathematics [10]. Several statistical models have been used to analyse students' performances in Mathematics in different levels of education for the purpose of improving their performances in Mathematics. Some of them are: ARIMA models which were used by Mussa [17] to forecast students' performances in Mathematics in certificate of secondary education examination in Zanzibar. Simple linear regression models were used by Lusana [12] to model pupils' academic performances in primary schools. In his study, Mathematics and other two subjects were analysed. Not only that but also, correlation model and binary logistic regression model were used by Mazana and colleagues [14] to investigate students' attitude towards learning Mathematics.

Despite all the recommendations proposed from these studies, students' performance in Mathematics have consistently remained poor. Therefore, the current researcher sees the need and necessity of conducting another study by using different methodology to contribute the effort of improving academic performances of the students in Mathematics. Therefore, to predict students' performances in Mathematics into different categories, the study of Predictive Discriminant Analysis (PDA) is required [4]. Therefore, the objective of this study is to develop the classification model that will be used to predict students' performances in Mathematics into three different categories of performances on the basis of the individual academic performances when he/she was in form five.

The term discriminant analysis first appeared in 1936 in works of R. A. Fisher in the article "The Use of Multiple Measurement in Taxonomic Problems" [16]. During the time between 1950s and 1960s, educational methodologists at Harvard University began to employ the PDA to do classification in academic issues [6]. Notable areas in education where PDA is successfully and frequently used are: prediction of academic performance of students, prediction of students' placement, prediction of students' dropout and prediction of students' to graduate [9].

Erimafa, Iduseri, and Edokpa [3] conducted a study to predict class of degree obtainable in a university system in Nigeria. They used linear discriminant analysis to predict students' class degree that they will achieve during the graduation. The data for this study were from students'

academic records for 100 levels and 200 levels, in the department of statistics. The model successfully classified 87.5% of the graduating students' class of degrees.

Divjak and Oreski [2] conducted a study at the Faculty of Organization and Informatics, University of Zagreb Croatia to predict students' academic performances by using discriminant analysis. In their study 113 students were involved and categorized into goal-oriented, learning oriented and relationship-oriented groups. According to the findings of the study, 54% of students were goal-oriented, 19 % were learning-oriented, and 21% were relationship-oriented.

Thomas [19] conducted a study to predict students' college completion intentions by using discriminant analysis. He developed a classification model to predict students' college completion intentions. Students with low, medium and high levels of college completion intention were involved during the study. In his study, 262 undergraduate students from 4 universities in Philippines were used to fill questionnaires. The result from findings showed that 26% of the respondents were assigned to the low group, 32% to the Medium group, and 42% to the high group by considering their completion intentions.

Therefore, many studies with different authors such as: Erimafa, Iduseri, and Edokpa [3], Divjak and Oreski [2], as well as Thomas [19], have shown how successively PDA was used to do prediction in different issues in education. In the reviewed literature above, there is no known or more recent study of PDA related to the problem of poor performance in Mathematics which is conducted in Tanzania. Therefore, the researcher sees the need and necessity of conducting the study of PDA to develop a classification model to be used for prediction, to predict students' performances in Mathematics, particularly in advanced secondary schools owned by the government in Iringa Rural District. Once a classification model is developed and applied, it will allow students, teachers, government and other educational stakeholders to understand in advance how well each student is expected to perform in the upcoming examinations, and giving them a chance to adopt appropriate measures as early as possible to improve students' performances in Mathematics.

## 2. Predictive Discriminant Analysis

Predictive Discriminant Analysis (PDA) is a multivariate technique that is used to predict group membership of observations into non-overlapping groups [6]. Basically, the study of PDA focus on developing a classification model and assessing its performance by using a statistical quantity called classification accuracy [7]. The classification model has two parts, that is, discriminant functions and classification rule. A set of discriminant functions is used to discriminate multidimensional observations coming from different groups in an optimal way. After that, the classification rule is used to assign observations into one of the possible non-overlapping groups

considered during the study [1].

Therefore, in this study four major tasks are to be completed in order to have a classification model to be used for prediction. That is: grouping observations, developing a set of discriminant functions, classifying observations by using a classification rule and validating the classification model. Therefore, when the classification accuracy achieved by the classification model is satisfactory, then the classification model can be used for prediction to predict students' performances in Mathematics into the three proposed groups.

## 3. Methodology

Data collected for this study, were quantitative secondary data of students' performances in Mathematics, who enrolled to study a two-year advanced secondary education from 2020 to 2022 in advanced secondary schools owned by the government in Iringa Rural District. From the students' academic records, scores of 126 students of the first midterm test, terminal examination, second midterm test and annual examination when they were in form five in the year of study in 2020/2021 were collected. From them, three groups/levels of students' performances in Mathematics each containing 42 observations were created by considering students' average scores and the grading system used by National Examinations Council of Tanzania (NECTA) to rank students' performances in advance secondary education. There are 7 grades which are: F(0-34), S(35-39), E(40-49), D(50-59), C(60-69), B(70-79) and A(80-100). Whereas F is considered as FAIL and all of the remaining grades are considered as PASS. The average scores of D(50-59), C(60-69), B(70-79) and A(80-100) were considered as high performance, S(35-39) and E(40-49) as medium performance and F(0-35) as poor performance. In this study, "average score" and "difference score" of students' performances in Mathematics were considered as the predictor variables which were linearly transformed from the first midterm test, terminal examination, second midterm test and annual examination which students sat for when they were in form five in the year of study in 2020/2021.

3.1. **Discriminant Functions.** To build discriminant functions, we need to find the linear combinations of predictor variables, that is, "average score" and "difference score" which maximize the difference between the three groups under the study, with the objective of establishing mathematical models that are able to discriminate between High, Medium and Low levels of students' performances in Mathematics with minimal error [8]. A set of linear discriminant functions that maximize the Fisher criterion, $J(W)$ is given by the equation:

$$Y = W^T X. \tag{1}$$

Where: $Y$ is a set of projection axes and $W$ is a projection matrix.

Basically LDA aims to find a projection matrix, $W$ which maximizes the Fisher criterion, $J(W)$ after projection [5]. When the ratio reaches its maximum value, observations within each group have the least amount of scatter and the groups become more separated from one another [11]. The Fisher criterion, $J(W)$ is given by:

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}. \tag{2}$$

Where: $S_W$ is within-class scatter matrix, while $S_B$ is between-class scatter matrix.

According to Li and Wang [11], the projection vector that has the highest Eigen value, has higher discrimination power between the groups. Therefore, to obtain $W$, the largest non-zero characteristic root (or Eigen value) $\lambda$ of $S_W^{-1} S_B$ is computed by using the following equation:

$$|S_W^{-1} S_B - \lambda I| = 0. \tag{3}$$

Hence, when $W$ that associated with $\lambda$ is computed, we get a set of two discriminant functions which are mutually uncorrelated. These two functions can be used to discriminate three groups of students' performances in Mathematics by creating boundaries between the groups, that is High, Medium and Low performances.

3.2. **Classification Rule.** In this study, the maximum posterior probability rule was used as a classification rule to assign Mathematics performances of students into High, Medium and Low performances. The maximum probability rule makes use of posterior probabilities of the group memberships that minimizes the total number of misclassification errors [6]. That is, an observation is assigned to the group for which the posterior probability is maximum [6].

*Posterior Probabilities.* Let $\hat{P}(j/x)$ be estimated posterior probability that, an observation belongs to group $j$ when $x$ value is observed. That is:

$$\hat{P}(j/x) = \frac{\hat{q}_j * \hat{f}(x/j)}{\sum_{j'=1}^{j} \hat{q}_{j'} * \hat{f}(x/j')}. \tag{4}$$

Where:

$\hat{f}(x/j)$ is an estimate of the conditional probability density function of $x$ given that observed value $x$ comes from group $j$.

$\hat{f}(x/j')$ is an estimate of the conditional probability density function of $x$ given that observed value $x$ does not comes from group $j$.

$\hat{q}_j$ is an estimated prior probability of an observation, $x$ that belongs to group $j$.

$\hat{q}_{j'}$ is an estimated prior probability of an observation, $x$ that does not belong to group $j$.

*Assignment Rule.* An observation with observed value, $x$ is assigned to group $j$ if $\hat{P}(j/x) > \hat{P}(j'/x)$ for $j \neq j'$. This implies an observation with observed value, $x$ is assigned to the group for which the value of posterior probability is maximum.

3.3. **Confusion Matrix.** In this study, a three ways confusion matrix was used to summarize the classification counts. Table 1 shows a three ways confusion matrix which summarizes students' performances in Mathematics which were classified by the classification model.

TABLE 1. **Three Ways Confusion Matrixt**

| | **Predicted Performance** | | | |
|---|---|---|---|---|
| **Actual Performances** | High | Medium | Low | **Total** |
| High | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_1$ |
| Medium | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_2$ |
| Low | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_3$ |
| **Total** | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | **n** |

Where; $n$ is a total number of observations of a given dataset, $n_1$, $n_2$ and $n_3$ are the number of observations in High, Medium and Low performances respectively before prediction. While $n_{.1}$, $n_{.2}$ and $n_{.3}$ are the number of observations in High, Medium and Low performances respectively after prediction. From Table 1 above, the main diagonal contains number of observations which are correctly classified by the model, and the remaining off-diagonal observations are the one which are wrongly classified by the model.

3.4. **Classification Accuracy.** In this study, the classification accuracy was used to evaluate the classification performance achieved by the classification model to classify students' performances in Mathematics into three proposed group. This statistic was computed by measuring the percentage of the proportion of correctly classified observations. That is:

$$Classification\ Accuracy = \frac{Number\ of\ correctly\ classified\ observations}{Total\ number\ of\ observations} \times 100\%. \quad (5)$$

$$= \frac{n_{11} + n_{22} + n_{33}}{n_1 + n_2 + n_3} \times 100\% \quad (6)$$

Therefore, when the value of classification accuracy is largely close to 100%, it indicates that, the classification model is valid.

## 4. RESULTS AND DISCUSSIONS

The analysis was done by using R programming software, and a seed of 66 was used during data partitioning to create training and test datasets. 60% of 126 observations were randomly selected

as training dataset to develop the classification model, that is, a set of discriminant functions and classification rule. While the remaining 40% of 126 observations were used as test dataset for validation of the classification model. Table 2 shows standardized coefficients and proportion of traces of linear discriminant functions. While Table 3 and 4 shows the confusion matrices for the training and test dataset respectively.

TABLE 2. **Standardized Coefficients and Proportion of Traces of Linear Discriminant Functions**

| | LD1 | LD2 |
|---|---|---|
| Average score | -0.2167 | -0.0090 |
| Difference score | -0.0057 | -0.1356 |
| Proportion of Trace | 0.9983 | 0.0017 |

TABLE 3. **Confusion Matrix of the Students' Performances in the Training Dataset**

| | **Predicted Performance** | | | |
|---|---|---|---|---|
| **Actual Performance** | High | Medium | Low | **Total** |
| High | 24 | 2 | 0 | 26 |
| Medium | 0 | 24 | 0 | 24 |
| Low | 0 | 0 | 25 | 25 |
| **Total** | 25 | 26 | 25 | **75** |

TABLE 4. **Confusion Matrix of the Students' Performances in the Test Dataset**

| | **Predicted Performance** | | | |
|---|---|---|---|---|
| **Actual Performance** | High | Medium | Low | **Total** |
| High | 15 | 1 | 0 | 16 |
| Medium | 0 | 18 | 0 | 18 |
| Low | 0 | 1 | 16 | 17 |
| **Total** | 15 | 20 | 16 | **51** |

Table 2 gives the standardized coefficients of the first and second linear discriminant function, which enables us to get a set of two straight lines which discriminate between High, Medium and Low performances of students in Mathematics by creating the boundaries between them. These lines help us to get a clear separation between the three groups of students' performances in Mathematics, even though it is not possible to identify which group is on which side of the lines. Therefore, when the classification rule is used, we can be able to identify which group is on which side of the lines, by assigning each student's performance in Mathematics into one of the three

groups under the study. Therefore, the first linear discriminant function, LD1 and the second linear discriminant function, LD2 for this study are given as:

$$LD1 = -0.2167 \times average\_score - 0.0057 \times difference\_score \quad (7)$$

$$LD2 = -0.0090 \times average\_score - 0.1356 \times difference\_score \quad (8)$$

The percentage of separation achieved by the first and the second linear discriminant function are 99.83% and 0.17% respectively. This means the LD1 explains 99.83% amount of variations in the training dataset to separate High, Medium and Low performances of students in Mathematics. While the LD2 explains only 0.17% amount of variations to separate High, Medium and Low performances of students in Mathematics. Therefore, to achieve a maximum discrimination of the students' performances in Mathematics, the LD1 and LD2 should be together as a set to discriminate students' performances in Mathematics.

Table 3 shows majority of students' performances in the training dataset were correctly classified except 2 students' performances which were High performances and wrongly classified as Medium performances. In High performance group, 24 out of 26 students' performances were correctly classified as High performances while 2 students' performances were wrongly classified as Medium performances. Hence, the model achieved the classification accuracy of 92.31% to classify students' performances in High performance group. Furthermore, from the table, it looks clearly that, no students' performances belong to either Medium or Low performance were wrongly classified into another group. This implies, the model achieved the classification accuracy of 100% to classify students' performances in Medium and Low performance groups. Therefore, a total of 73 out of 75 students' performances in mathematics in the training dataset where correctly classified by the model. This implies, the overall classification accuracy achieved by the model to classify students' performances in Mathematics in the training dataset is 97.33%.

Table 4 shows majority of students' performances in the test dataset were correctly classified except 2 students' performances (one from High performance group and another from Low performance group) were wrongly classified into Medium performance group. In High performance group, 15 out of 16 students' performances were correctly classified as High performance while 1 student's performance was wrongly classified as Medium performance. Hence, the classification accuracy achieved by the classification model to classify students' performances in High performance is 93.75%. In the Medium performance, all 18 students' performances were correctly classified as Medium performance and hence the model achieves the classification accuracy of 100% in this group. Furthermore, in Low performance, 16 out of 17 students' performances were correctly classified as Low performance while 1 student's academic performance is wrongly classified as Medium performance. Hence the classification accuracy achieved by the classification model to classify

students' performances in High performance group was 94.12%. Therefore, a total of 49 out of 51 students' performances in Mathematics in the test dataset where correctly classified by the model. This implies, the overall classification accuracy achieved by the classification model to classify students' performances in Mathematics in the test dataset was 96.08%. Since the overall value of the classification accuracy is largely close to 100%, it implies that, the classification model is reliable and convenient to be used for prediction to predict students' performances in mathematics.

## 5. Conclusions

The focus of the discussion of PDA in this study concentrated on developing a classification model to be used for prediction, to predict students' performances in Mathematics into three different levels of academic performances. Based on the two overall values of classification accuracy computed from the training and test dataset, this study suggests, the classification model is valid and reliable to be used for prediction. Therefore, once we have a set of new data of students' performances in Mathematics of the first midterm test, terminal examination, second midterm test and annual examination when they were in form five and transforming them into "average score" and "difference score", then the classification model can be used to predict performances of those new students into one of the three groups of performances (High, Medium and Low) under the study. Once prediction is done, educational stakeholders particularly Mathematics teachers will be able to identify poor performers and hence helping them to improve their performances in Mathematics in the upcoming examinations.

## References

[1] K. E. Ahmad, Z. F. Jaheen, A. A. Modhesh, Estimation of a discriminant function based on small sample size from a mixture of two Gumbel distributions. Comm. Stat. Simul. Comp. 39 (2010) 713–725. https://doi.org/10.1080/03610911003624867.

[2] B. Divjak, D. Oreski, Prediction of academic performance using discriminant analysis. Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, (2009) 225–230. https://doi.org/10.1109/ITI.2009.5196084.

[3] J. T. Erimafa, A. Iduseri, I. W. Edokpa, Application of discriminant analysis to predict the class of degree for graduating students in a university system. Int. J. Phys. Sci. 4 (2009) 16–21.

[4] F. Faulkner, A. Hannigan, O. Fitzmaurice, The role of prior mathematical experience in predicting mathematics performance in higher education. Int. J. Math. Educ. Sci. Technol. 45 (2014) 648–667. https://doi.org/10.1080/0020739X.2013.868539.

[5] Q. Gu, Z. Li, J. Han, Linear discriminant dimensionality reduction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, (2011), 549–564.

[6] C. J. Huberty, Why multivariable analyses? Educ. Psych. Measure. 54 (1994) 620–627. https://doi.org/10.1177/0013164494054003005.

[7] C. J. Huberty, R. M. Barton, An introduction to discriminant analysis. Measure. Eval. Counsel. Develop. 22 (1989) 158–168. https://doi.org/10.1080/07481756.1989.12022925.

[8] C. J. Huberty, S. Olejnik, Applied MANOVA and discriminant analysis (2nd ed., Vol. 498). John Wiley & Sons. (2006).

[9] A. Iduseria, J. E. Osemwenkhaeb, On the use of predictive discriminant analysis in academic prediction. J. Nigerian Stat. Assoc. 29 (2017) 71–80.

[10] K. Jayarajah, R. M. Saat, R. A. A. Rauf, A review of science, technology, engineering & mathematics (STEM) education research from 1999-2013: A Malaysian perspective. Eurasia J. Math. Sci. Techn. Educ. 10 (2014) 155–163. https://doi.org/10.12973/eurasia.2014.1072a.

[11] C. Li, B. Wang, Fisher linear discriminant analysis. CCIS Northeastern University. (2014).

[12] A. M. Lusana, Modelling of primary school pupils' academic performance in Tanzania. The University of Dodoma. (2018).

[13] M. Y. Mazana, C. S. Montero, R. O. Casmir, Assessing students' performance in mathematics in Tanzania: the teacher's perspective. Int. Elect. J. Math. Educ. 15 (2020) em0589. https://doi.org/10.29333/iejme/7994.

[14] M.Y. Mazana, C.S. Montero, R.O. Casmir, Investigating Students' Attitude towards Learning Mathematics. Int. Elect. J. Math. Educ. 14 (2018), 207-231. https://doi.org/10.29333/iejme/3997.

[15] I. M. Mbiti, The need for accountability in education in developing countries. J. Econ. Perspect. 30 (2016) 109–132. https/doi.org/10.1257/jep.30.3.109.

[16] M. Misankova, K. Kocisova, Strategic implementation as a part of strategic management. Procedia-Soc. Behav. Sci. 110 (2014) 861–870. https://doi.org/10.1016/j.sbspro.2013.12.931.

[17] A. M. Mussa, Forecasting performance of students in mathematics in certificate secondary education examination (CSEE) in Zanzibar. The University of Dodoma. (2017).

[18] D. Suryadarma, A. Suryahadi, S. Sumarto, F. H. Rogers, Improving student performance in public primary schools in developing countries: Evidence from Indonesia. Education Economics, 14 (2006) 401–429. https://doi.org/10.1080/09645290600854110.

[19] D. Thomas, Predicting student college completion intention: A discriminant analysis. ASEAN J. Manage. Innov. 1 (2014) 41–55.